NOISE REDUCTION FOR SPEECH ENHANCEMENT USING NON-LINEAR WAVELET PROCESSING

THESIS

Hassan Dehmani
First Lieutenant, Royal Moroccan Air Force (RMAF)

AFIT/GCE/ENC/94J-1

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GCE/ENC/94J-1

NOISE REDUCTION FOR SPEECH ENHANCEMENT USING NON-LINEAR WAVELET PROCESSING

THESIS

Hassan Dehmani
First Lieutenant, Royal Moroccan Air Force (RMAF)

AFIT/GCE/ENC/94J-1

AFIT/GCE/ENC/94J-1

NOISE REDUCTION FOR SPEECH ENHANCEMENT USING NON-LINEAR WAVELET PROCESSING

THESIS

Presented to the Faculty of the School of Engineering

of the Air Force Institute of Technology

Air University

In Partial Fulfillment of the

Requirements for the Degree of

Master of Science in Computer Engineering

Hassan Dehmani, B.S.E.E

First Lieutenant, Royal Moroccan Air Force (RMAF)

June, 1994

Preface

I would like to thank all the monkeys and elephants of Africa for making this thesis a very successful one. Also I would like to thank Taco Bell, McDonald's, and Burger-King for providing flexible hours and food service. Finally, I would like to thank my "mistress", Maddie, for those romantic nights in room 2000 and Sparc and Sparc2 for being there whenever I needed them.

This work is dedicated to all my loved ones, especially my family who kept me alive by those periodic humorous phone calls from the mother land (Morocco-Africa). Thanks to my dad, Mohammed, my mom, Fatna, my brothers, Mostafa, Rachid, Noaman, and the cute little Amine, and my sisters, Ilham, Nazha, and Wafaa. Their love, prayers, and moral support made this work possible.

I would like also to thank my advisor, Maj. Gregory T. Warhola, PhD. His support, encouragements, insights, and guidance were a major factor in the success of this thesis. You will have an exceptional place in both my heart and my mind. Your dedication to excellence is a virtue only special people can have. Thanks to Capt. Joseph Sacchini, Dr. Norman Weyrich (Germany), and Dr. Bruce Suter for providing useful suggestions in this thesis.

Laura Suzuki, all I can say is this: your help, time, dedication, encouragements, support, and patience impressed me considerably (if you don't go to heaven, something is wrong with the concept of life). Thanks to John Colombi for being there at all times. Also thanks to Kabrisky, Desimio, Capt. Bruce Anderson and all the instructors here at AFIT, you are the best. Thanks to Capt. Armin Sayson and wife Malu (Philippines) and Felix (Taiwan) for the fun we had at Georgio's and the food and donuts we ate to make it through AFIT's holy war. The "We" team is the best concept on planet earth.

Finally, thanks to my lovely wife Donia for cooking, cleaning, the rides to AFIT, the moral support, the massages after each long night at AFIT, and most importantly, the lovely picnics we had at the lab (room 2000) while the computer was running. Your support is part of my success.

Thanks for being domestic. Remember, behind each successful man there is a woman, and behind each successful officer there is a leader at home: it is the spouse. I love you!.

Hassan Dehmani

Table of Contents

List of Figures

List of Tables

Abstract

The problem of speech enhancement presents many obstacles in the speech processing field. This thesis develops several speech de-noising systems (SDS) that can be used in the time, Fourier, and the wavelet domains. We present two different thresholding techniques, the soft thresholding technique (STT) and the hard thresholding technique (HTT). The application of these thresholding techniques to noisy speech data is discussed. The combination of both the Fourier and wavelet domains in speech de-noising proves to yield the best results in terms of speech intelligibility. Informal listening tests are conducted in order to compare the effects of using the STT, the HTT, the noisy phase, the time domain, the Fourier domain, and the wavelet domain.

NOISE REDUCTION FOR SPEECH ENHANCEMENT USING NON-LINEAR WAVELET PROCESSING

I. Introduction

1.1 Background

In recent years, many speech processing scholars have developed speech systems that have some degree of success when used with speech data acquired under near-ideal conditions. By far the majority of recognition and encoding schemes have been developed and tested using speech recorded on very sophisticated equipment in a quiet environment. As speech processing has moved from the ideal laboratory conditions to the field, it has become significantly important to face the problems imposed by the presence of ambient noise. Once in the real world, most of the speech processing systems, especially speech recognition and speech encoding systems, fall very short on their promises. Speech degraded by ambient noise has most of its formant's structure detectable by the human listener, however, the human listener cannot listen to speech under degraded conditions for a long time without suffering auditory fatigue (19). In order to reduce the effects of ambient noise, many techniques for enhancement of noisy speech have been developed.

The main objective of the speech enhancement is to attenuate the intensity of the noise, while preserving the overall structure (i.e., pitch, formants, etc.) and intelligibility of speech. In particular, the military environment is one of the most crucial environments where speech data is vulnerable to ambient noise, especially noise due to the engines of tanks, military vehicles, helicopters, airplanes, and others.

1.2 Problem Statement

The problem considered in this thesis is to enhance noisy speech data and still preserve intelligibility. In order to accomplish this goal, we propose to develop a speech processing scheme using both wavelets and the thresholding techniques.

The United States military is carrying on intensive research in order to develop systems that are very reliable and very robust in enhancing speech data degraded by ambient noise. One of the new areas of this research is the use of wavelets in order to explore their unique filtering abilities with noisy speech data. In the last decade, the theory of wavelets has grown significantly, and has promised to change both signal and information processing. The major advantage of wavelets over the classic signal processing tools (i.e., Fourier transform), is their unique ability to decompose a signal into orthogonal resolution levels. This unique property, makes wavelets one of the best tools to use with signals composed of many high energy peaks of frequencies, such as speech.

In general, noise is a broad-band signal. The ability of wavelets to decompose a signal into various bands of frequencies, allows us to locate noise at certain frequency bands and eliminate it, however, at the expense of affecting the formants structure of the signal degraded by this noise. In order to avoid the distortion of the underlying signal, we resort to the use of many thresholding techniques which are based on the general statistics of the ambient noise. Hard thresholding is a technique that eliminates all data samples below a fixed threshold in absolute value. On the other hand, soft thresholding is a technique that eliminates all data samples below a fixed threshold in absolute value, and pulls towards zero all data above the threshold, by the amount of the threshold in absolute value. The use of thresholding helps decrease the amount of noise, while preserving most of formants' structure of the underlying signal.

1.3 Scope

This thesis is limited to the development of different speech de-noising systems to process speech to which various amounts of white Gaussian noise have been added (signal-to-noise ratios vary between -10db to 10db). These systems are based on the use of wavelets, Fourier, and non-linear statistical processing of speech data from the TIMIT data base. Quantitative squared error criteria and qualitative listening tests are performed.

No attempt to automatically determine pitch, silent, voiced, or unvoiced portions is made. These are assumed to be known. The algorithm developed is intended to be one subsystem of a pre-processor used to remove noise from noisy speech before use by other speech processing systems (e.g., speech identification, speech recognition, etc) or by human listeners.

The necessary mathematical background in wavelets, Fourier, and non-linear statistical methods, which are necessary to understand the de-noising systems developed in this thesis is presented.

1.4 Approach

The noisy speech signal is decomposed into voiced, unvoiced, and silent portions. The silent portions are used to estimate the variance of the noise which is assumed to be white Gaussian noise. The voiced portions are subjected to the thresholding techniques. Depending on the method used, we may process speech in time, frequency (Fourier), wavelet, or any combination of these three domains. The phase of the noisy voiced speech may be saved before processing the noisy voiced speech. On the other hand, both the unvoiced and silent portions are multiplied by a ratio to be discussed later. Before processing any speech segment, each portion (i.e., voiced, unvoiced, and silent) is multiplied by a window function to be defined later.

1.5 Objectives

The objectives of this research are to answer the following four questions:

1. Can we enhance noisy speech by applying both wavelets and the thresholding techniques?
2. Under what conditions do the application of wavelets and the thresholding techniques to noisy speech data yield intelligible results?
3. Can we use both wavelets and Fourier analysis to enhance noisy speech?
4. How do wavelets and the thresholding techniques affect the quality of the de-noised speech?

1.6 Equipment and Materials

The following tools were crucial to this research:

1. SPARC 2 workstations is used for coding and testing purposes.
2. ANSI C is the programming language for all codes developed for this research.
3. Mathematica is used for developing graphs and bar-charts.
4. ESPS-4 (Entropic Signal Processing System) is used for both spectrograms and listening tests.
5. LaTeX is used to typeset this document.
6. TIMIT data base.

1.7 Organization

In chapter two, we present past and current research in the area of enhancement of noisy speech. In chapter three, we discuss the necessary wavelet, Fourier, and thresholding theories. Based on the results and theories of chapter three, we present, in chapter four, eight de-noising systems. In chapter five, we test the de-noising systems of chapter four with actual noisy speech data and analyze the results in terms of both error and spectrogram analysis as well as informal listening tests. Finally, in chapter six, we present the thesis conclusions and recommendations.

II. Literature Review

2.1 Introduction

This chapter focuses on evaluating past techniques and research in the area of enhancing noisy speech. These techniques cover several methods used to solve the problem of eliminating some of the noise from a speech signal. Because of the similarities between the different techniques, we present each method in chronological order in order to understand some of the problems encountered in the field of speech processing.

2.2 Recent Developments In Enhancing Noisy Speech

Enhancing noisy speech presents three major problems:

a. detecting the presence of noise.

b. estimating the noise power.

c. differentiating between speech and non-speech signals.

The quality and intelligibility of the resulting speech signal depend on the method used and on the assumptions made to locate and estimate the noise.

2.2.1 Suppression of Acoustic Noise In Speech Using Spectral Subtraction. In 1979, Steven Boll presented a simple technique (Spectral Subtraction) to enhance speech degraded by additive white noise (3). His technique (among the best techniques during the early eighties) is well known in the speech processing field. His algorithm measures the signal present during non-speech activity and use it as an estimate of the noise. The spectrum of the estimated noise is then subtracted from that of the noisy speech . If we assume that speech is a stationary signal and that the noise is additive and uncorrelated, then we can present the noisy speech signal as

$$y(t) = s(t) + n(t), \tag{2.1}$$

where s and n are the speech and noise signals, respectively, where both are real. Taking the Fourier transform (see equation 3.137) of equation 2.1, we obtain

$$\tilde{y}(\omega) = \tilde{s}(\omega) + \tilde{n}(\omega). \tag{2.2}$$

The power of the above spectra is given by

$$|\tilde{y}(\omega)|^2 = |\tilde{s}(\omega)|^2 + |\tilde{n}(\omega)|^2 + 2[\mathbf{Re}[\tilde{s}(\omega)]\mathbf{Re}[\tilde{n}(\omega)] + \mathbf{Im}[\tilde{s}(\omega)]\mathbf{Im}[\tilde{n}(\omega)]]. \tag{2.3}$$

Since the noise and signal random variables are assumed to be uncorrelated, the expected value (see equation 3.7) of the crossproduct terms of equation 2.3 are eliminated and the expected power spectra can then be related by (19)

$$|\tilde{y}(\omega)|^2 = |\tilde{s}_e(\omega)|^2 + |\tilde{n}_e(\omega)|^2, \tag{2.4}$$

where $|\tilde{n}_e(\omega)|^2$ and $|\tilde{s}_e(\omega)|^2$ are estimates of the noise and speech powers, respectively.

If we can obtain a satisfactory estimate of $|\tilde{n}(\omega)|^2$, we can recover $|\tilde{s}(\omega)|^2$ by using equation 2.4, since we know the power $|\tilde{y}(\omega)|^2$. In practice, the noise is estimated by observing the signal during non-speech activity (19). The result is

$$|\tilde{s}_e(\omega)|^2 = |\tilde{y}(\omega)|^2 - |\tilde{n}_e(\omega)|^2. \tag{2.5}$$

Using the results from equation 2.5, Boll subtracted the magnitude spectra themselves instead of the power spectra, and since the magnitude is a positive quantity, any negative output is set to zero (19). The above process can be viewed as a filtering operation defined by

$$|\tilde{s}_e(\omega)| \quad = \quad |\tilde{y}(\omega)| - |\tilde{n}_e(\omega)|$$

$$= |\tilde{y}(\omega)| \left(1 - \frac{|\tilde{n}_e(\omega)|}{|\tilde{y}(\omega)|}\right)$$
$$= |\tilde{y}(\omega)|\tilde{h}(\omega), \tag{2.6}$$

where the filter $\tilde{h}$ is given by

$$\tilde{h}(\omega) = \left(1 - \frac{|\tilde{n}_e(\omega)|}{|\tilde{y}(\omega)|}\right) \tag{2.7}$$

where $0 \leq |\tilde{h}(\omega)| \leq 1$. Since the negative amplitudes are not allowed, Boll used the filter $\tilde{h}$ to define a half-wave rectification filter $\tilde{h}_R$ as

$$\tilde{h}_R(\omega) = \frac{\tilde{h}(\omega) + |\tilde{h}(\omega)|}{2}. \tag{2.8}$$

In order to recover the estimated speech signal $s_e(t)$, we need to take the inverse Fourier transform (see figure 2.1). However, we need the phase of $\tilde{s}_e(\omega)$. Boll approximated this phase by the phase of the known noisy signal $\tilde{y}(\omega)$. The recovered signal can then be obtained using the following equation

$$\tilde{s}_e(\omega) = |\tilde{s}_e(\omega)|e^{i\theta}, \tag{2.9}$$

where θ is the phase of $\tilde{y}(\omega)$.

In order to account for the case where the speech is absent, Boll modified his algorithm to allow a second pass to further reduce the residual noise left after the application of the spectral subtraction. The residual noise can be further attenuated without distorting the speech signal(3).

2.2.2 Speech Enhancement By Fourier-Bessel Coefficients Of Speech And Noise. In 1990, F.S. Gurgen and C.S. Chen introduced an enhancement technique for noisy speech based on the Fourier-bessel (FB) expansion of the speech and noise (11). The method is based on the subtraction

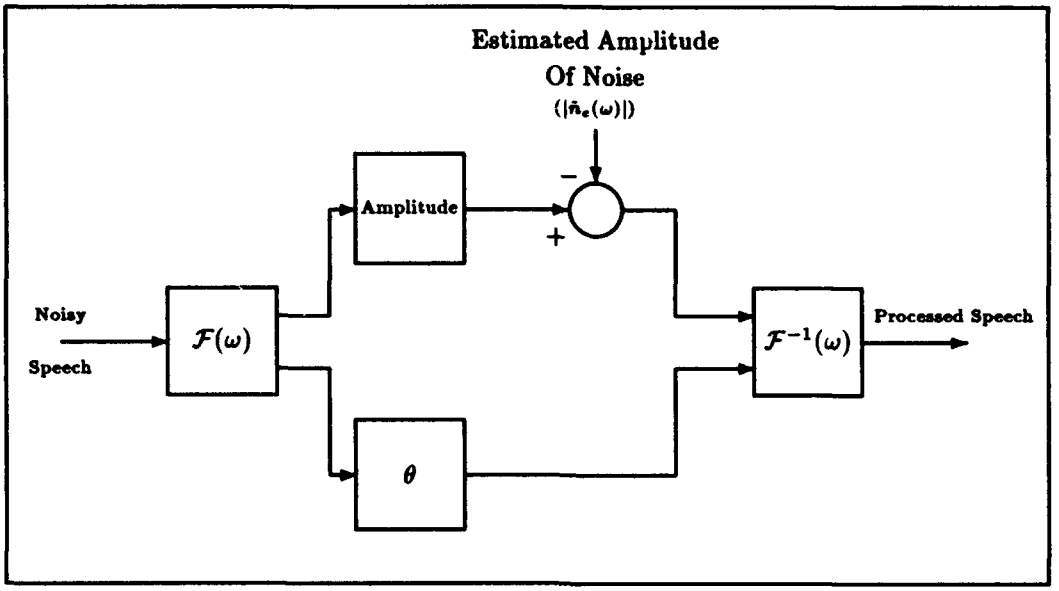

Figure 2.1 Spectral Subtraction By Steven Boll

of the FB coefficients of the estimated noise from the coefficients of the noisy speech. The difference in two sets of coefficients is then used to synthesize the enhanced speech .

2.2.2.1 Spectral Properties of Fourier-Bessel Coefficients. The solution of the wave equation inside cylindrical structures (tubes) includes the first kind of the Bessel function (22). In their method, Gurgen and Chen model the vocal tract as a cylindrical tube. The speech signal is represented using the first kind and first order Bessel functions, $J_1(t)$, as the basis functions for expansion. This representation is called a Fourier-Bessel (FB) expansion.

The FB expansion of the speech signal is achieved by using $J_1(\alpha_m t)$ as basis functions of representation, where $\alpha_m = \frac{t_m}{A}$, t_m is the m^{th} root of $J_1(t) = 0$, and A is the duration of the time frame under analysis. The decomposition describes a speech signal as a linear combination of the orthogonal basis functions

$$s(t) = \sum_{m=1}^{\infty} c_m J_1(\alpha_m t). \tag{2.10}$$

The set $\{J_1(\alpha_m t)\}$ is orthogonal with respect to the weighting function t, and the c_m coefficients in equation 2.10 are given by

$$c_m = \frac{2}{[A^2(J_0(t_m))^2]} \int_0^A t\, s(t) J_1(\alpha_m t)\, dt. \tag{2.11}$$

By taking the Fourier Transform of $J_1(\alpha_m t)$, Gurgen and Chen showed that the FB series behaves like a low-pass filter. By using the magnitude and the phase spectrum, it is possible to calculate the maximum frequency achieved with the number of the roots of $J_1(\alpha_m t)$ as (11)

$$f_{max} = \frac{t_m}{2\pi A}. \tag{2.12}$$

2.2.2.2 Noise suppression using FB expansion. Just like Boll's method, the speech signal $s(t)$ is assumed to be degraded by uncorrelated additive noise $n(t)$ where the noisy speech signal $y(t)$ is given by

$$y(t) = s(t) + n(t). \tag{2.13}$$

Taking the FB expansion of the above signal, we get

$$y_m = s_m + n_m, \tag{2.14}$$

where $m = 1, 2, 3, \ldots$.

Experimentally, Gurgen and Chen showed that the FB coefficients representation, with up to 150 coefficients and 10ms analysis frame , introduces a low-pass filtering effect on the speech signal by attenuating its high-frequency region. Therefore, the noise which is assumed to contain most of the high frequency components, can be suppressed by using an appropriate number of coefficients in the synthesis of the signal (11).

Since y_m is known (raw data), if we can obtain a satisfactory estimate of the noise level and calculate its FB expansion we can get an estimate of the enhanced speech signal as

$$s_m = y_m - n_m. \tag{2.15}$$

The estimation of the noise is based on two different techniques, the single-microphone case and the two-microphone case. In the single-microphone case, the noise estimate is accomplished by detecting the speech/non-speech intervals using energy thresholds to locate the silence intervals where the energy of the noise can be estimated. In the two-microphone case, a reference microphone path is used to estimate the noise and calculate its FB coefficients. A primary microphone path is used to calculate the FB coefficients of the noisy speech. The difference between these two paths is used to estimate the FB coefficients of the enhanced signal (11).

2.2.3 Adaptive Noise Reduction Using Discrimination Functions. Most speech enhancement techniques (e.g., spectral subtraction by S. Boll) are based on using speech detectors to locate the non-speech activities in a speech signal and use that information to estimate noise. The quality of the results depends heavily on the quality of the speech detectors used in the analysis. The Discrimination Function Minimization (DFM) method does not use a speech detector and does not assume stationarity of the noise over an entire speech period . The purpose of the DFM is to define a function that differentiates between clean and noisy speech signals in order to reduce the noise in the noisy speech signal (10). Based on essential features of speech and ambient noise, the DFM uses a single-microphone adaptive filtering approach and minimizes a mean square error function.

2.2.3.1 Discrimination Function Minimization (DFM). The DFM technique involves two steps:

1. Definition of a Discrimination Function $J(\mathbf{x})$

The discrimination function $J(\mathbf{x})$ is defined for a vector $\mathbf{x} = \{x_i\}_{\{0 \le i \le N-1\}}$ such that

$$J(\mathbf{x})|_{\mathbf{x} \in S \cap N} < J(\mathbf{x})|_{\mathbf{x} \in N}, \tag{2.16}$$

where

a. $S = \{s\}$, set of segments of clean speech.

b. $N = \{n\}$, set of segments of noise.

c. $S \cap N = \{y = s + n\}$, set of noisy speech segments.

The above equation states that the value of J for pure noise signals is greater than that of speech and noise signals.

2. Filtering or suppression of the noise based on setting the coefficients of the filter h such that J is minimized (see figure 2.2).

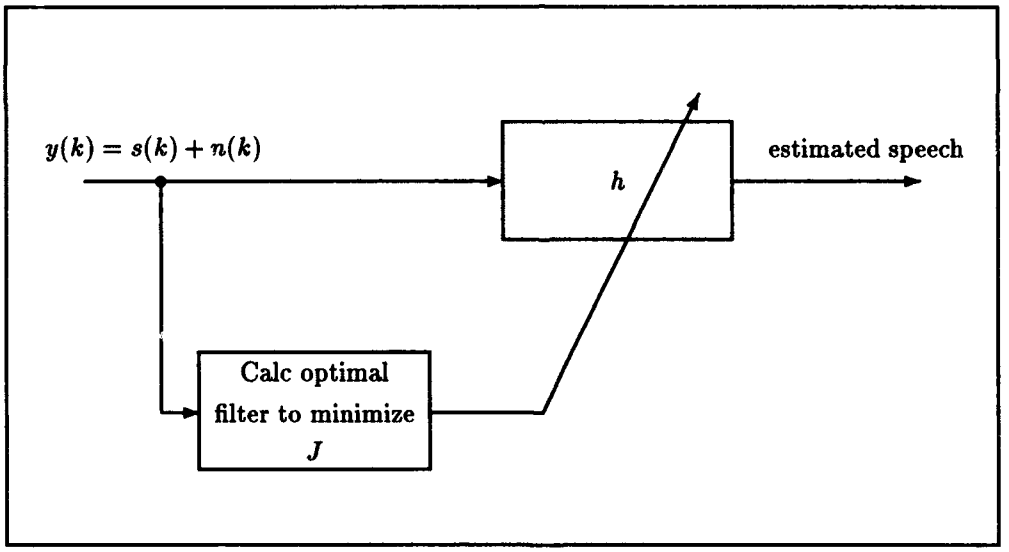

Figure 2.2 Block Diagram C The DFM Noise Reduction

2.2.3.2 Example Of A Discrimination Function $J_R(\mathbf{x})$. Since the rate of change of the noise parameters (e.g., autocorrelation) are less than those of speech signals, the authors concluded that it is possible to derive some discrimination functions directly from the dynamics of the speech sample-variance. Experimentally, the authors found that the rate of change,

$\Delta v(i)$, of a speech frame variances $\sigma(i)$ and the duration of the so called stationary periods, where $\Delta v(i) < \Delta v_{Thres}$ are two discriminating features that can be used to filter the noise out of a noisy speech signal (10).

Let s be a clean speech segment degraded by uncorrelated additive noise n, the noisy speech signal y is given by

$$y = s + n, \tag{2.17}$$

and define N_R be a subset of the noise set N such that, the length of each vector in N_R is longer than τ_{max}, the maximum length of a stationary period in a speech signal (i.e., $\tau_{max} = 200$ms). The discrimination function is then defined as

$$J_R(\mathbf{x})|_{\mathbf{x}\in S\cap N} < J_R(\mathbf{x})|_{\mathbf{x}\in N_R}, \tag{2.18}$$

where $N_R \subset N$. Let $J_R(\mathbf{x})$ be a discrimination function defined over a data frame of length N. The sample-variances are calculated for each sub-frame of length L. These sub-frames are non overlapping and, therefore, we have $p = \frac{N}{L}$ sub-frames.

The sample variance for each sub-frame is defined as

$$\sigma(i) = \sqrt{\frac{1}{L}\sum_{k=0}^{L-1} \hat{s}^2(iL-k)}, \tag{2.19}$$

where $i = 1, 2, \ldots p$ and $\hat{s}(j)$ is the filtered signal at time j, which is calculated using the input vector $\mathbf{y}$ and a transversal filter $\mathbf{h}$ of order M such that:

$$\hat{s}(j) = \sum_{n=0}^{M-1} h_n y(j-n), \tag{2.20}$$

for $j = 0, 1, ..N-1$. Now define the absolute value of the relative change of the variance as

$$\Delta v(i) = \left| \frac{\sigma(i) - \sigma(i-1)}{\sigma(i-1)} \right|, \tag{2.21}$$

and an exponential weighting factor as

$$w(i) = r^i, \tag{2.22}$$

where $0 < r \leq 1$, and $i = 1, 2, \ldots p$.

Using the above definitions, we can define two discrimination functions:

1. A first discrimination function that maximizes the relative changes of variance defined as

$$J_{R1}(\mathbf{x}) = \left(\sum_{k=0}^{p-1} w(k) \Delta v(p-k) \right). \tag{2.23}$$

2. A second discrimination function that minimizes the durations of the stationary periods under analysis defined as

$$J_{R2}(\mathbf{x}) = \sum_{k=0}^{p-1} w(k) e^2(p-k), \tag{2.24}$$

where $e(i)$ is defined as

$$e(i) = \begin{cases} \tau - \tau_{max} & \text{if } \tau \geq \tau_{max} \text{ and } \Delta v(i) < \Delta v_{Thres} \\ 0 & \text{otherwise,} \end{cases} \tag{2.25}$$

where for $i = 1, 2, \ldots p$, the value of $e(i)$ is the excess time beyond the frames period τ_{max}. The entire discrimination function can be defined as

$$J_R(\mathbf{x}) = c_1 J_{R1}(\mathbf{x}) + c_2 J_{R2}(\mathbf{x}), \tag{2.26}$$

where c_1 and c_2 are normalizing factors. The minimization of J_R in order to find the coefficients of the filter h has two consequences: J_{R1} maximizes the relative changes of the variances and, according to equation 2.25, J_{R2} minimizes the durations of the stationary periods (10).

The accuracy of the DFM method depends heavily on the validity of the discrimination function. Besides the fact that the DFM does not require a speech activity detector, the main advantage of the DFM is that the filter h adapts to the changes of the noise patterns throughout the speech signal.

2.2.4 Other Speech Enhancement Techniques. Many speech processing researchers model speech as a sum of sinusoidal periodic functions. Kobatake, Karou, and Sheng approached the speech enhancement problem by means of the maximum likelihood estimation (MLE). The authors segmented the speech signal into frames and sub-frames and then, by maximizing an *a posteriori* probability density function, they estimated the Fourier coefficients of the voiced portions at a specific frame (15).

In 1989, Nadeem A. Bashir, a graduate student at the Air Force Institute Of Technology (AFIT), developed a system in order to enhance the quality of mutilated speech. His technique analyses the damaged speech in the Fourier domain and then, based on known properties of normal speech, such as periodicity of voiced speech, a computer program generates a set of sinusoids whose amplitudes and phases are derived directly from the speech signal itself. These sinusoids are used to reconstruct a cleaner and clearer version of the mutilated speech (13).

III. Stein's Criteria, Wavelet, And Fourier Theory

3.1 Introduction

In this chapter, we present three main topics: Stein's criteria, wavelets, and Fourier analysis. Stein's criteria defines both the necessary conditions to estimate the mean of an independent normal random vector, as well as a method for estimating the risk associated with the mean estimation technique. Next, we present the necessary wavelet theory and how it can be related to Stein's criteria. In fact, we will prove that the wavelet coefficients of an independent normal random vector, are themselves independent and normal. This property of the wavelet coefficients makes them candidates to use with Stein's criteria. Finally, we present the Fourier transform and some of its properties and we will prove that the Fourier coefficients (with some restrictions to be discussed later) can be used with Stein's criteria.

Using the theory of Stein, we present two different thresholding techniques, the hard thresholding technique (HTT) and the soft thresholding technique (STT). These thresholding techniques will be used in our experiments dealing with de-noising speech. Throughout this chapter, we will assume that all random vectors are independent, normal, and have the same variance.

3.2 Stein's Unbiased Estimate Of Risk (SURE)

Given a normal random vector, $\vec{X} = (X_0, X_1, X_2, \ldots, X_{N-1})$, whose elements ,$X_i$, are **independent normal** random variables with arbitrary means and the **same** variance σ^2 such that for $i = 0, 1, 2, \ldots, N-1$

$$X_i \sim N(\mu_i, \sigma^2), \tag{3.1}$$

Charles Stein, a statistician at Stanford University, introduced a simple equation (SURE) to estimate the error associated with the estimation $\hat{\vec{\mu}} = (\hat{\mu}_0, \hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_{N-1})$ of the true mean,

$\vec{\mu} = (\mu_0, \mu_1, \mu_2, \ldots, \mu_{N-1})$, of the normal random vector $\vec{X}$ by

$$\hat{\vec{\mu}} = \vec{X} + g(\vec{X}), \tag{3.2}$$

where $g : \mathbf{R^N} \longrightarrow \mathbf{R^N}$ is an almost differentiable function to be defined later (20).

Stein's theory can be used with any normal random vector with independent random variables whose variances are identical. The next sections provide a detailed derivation of Stein's error equation which we will use with both wavelets and Fourier. Stein developed his criteria by first deriving the basic equations for a standard normal random variable (zero mean and variance of one) and then, he extended the results to the case of several arbitrary normal random variables with the same variance. It is important to understand that all the normal random variables are assumed to be independent and have the same variance with an arbitrary mean.

3.2.1 Standard Normal Distribution: $X \sim N(0,1)$. Let X be a real random variable with a standard normal distribution

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}. \tag{3.3}$$

The derivative of the above probability density function (pdf) is

$$\phi'(x) = -x\phi(x), \tag{3.4}$$

and let g be an indefinite integral of the Lebesgue measurable function g' such that

$$g : \mathbf{R} \longrightarrow \mathbf{R}, \tag{3.5}$$

and

$$\mathbf{E}\Big\{\big|g'(X)\big|\Big\} < \infty, \tag{3.6}$$

where **E** is the expectation operator defined by

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x\phi(x)\,dx, \tag{3.7}$$

and g' is the derivative of the function g. We shall show that

$$\mathbf{E}[g'(X)] = \mathbf{E}[X\,g(X)]. \tag{3.8}$$

First of all, we have the following identities concerning the standard normal distribution

$$\begin{aligned} \phi(x) &= \int_{-\infty}^{x} \phi'(z)\,dz \\ &= \int_{-\infty}^{x} -z\phi(z)\,dz. \end{aligned} \tag{3.9}$$

Since $\phi(x) = \phi(-x)$ and $\phi'(x) = -x\phi(x)$, we have the following relations

$$\begin{aligned} \phi'(-x) &= x\phi(-x) \\ &= x\phi(x) \\ &= -\phi'(x), \end{aligned} \tag{3.10}$$

and we can then write

$$\begin{aligned} \phi(x) &= \phi(-x) \\ &= \int_{-\infty}^{-x} \phi'(-z)\,dz \end{aligned}$$

$$
\begin{aligned}
&= \int_{-\infty}^{-x} -\phi'(z)\,dz \\
&= \int_{-\infty}^{-x} z\phi(z)\,dz \\
&= \int_{-\infty}^{-x} -z\phi(-z)d(-z) \\
&= \int_{x}^{\infty} z\phi(z)\,dz. \qquad (3.11)
\end{aligned}
$$

Using the above equalities, we get

$$
\begin{aligned}
\mathbf{E}[g'(X)] &= \int_{-\infty}^{\infty} g'(x)\phi(x)\,dx \\
&= \int_{-\infty}^{0} g'(x)\phi(x)\,dx + \int_{0}^{\infty} g'(x)\phi(x)\,dx \\
&= \int_{-\infty}^{0} g'(x) \int_{-\infty}^{x} -z\phi(z)\,dz\,dx + \int_{0}^{\infty} g'(x) \int_{x}^{\infty} z\phi(z)\,dz\,dx.
\end{aligned}
\qquad (3.12)
$$

Using Fubini's theorem (2), we can switch the order of integration and get

$$
\begin{aligned}
\mathbf{E}[g'(X)] &= -\int_{-\infty}^{0} z\phi(z) \int_{z}^{0} g'(x)\,dx\,dz + \int_{0}^{\infty} z\phi(z) \int_{0}^{z} g'(x)\,dx\,dz \\
&= \int_{-\infty}^{0} z\phi(z) \int_{0}^{z} g'(x)\,dx\,dz + \int_{0}^{\infty} z\phi(z) \int_{0}^{z} g'(x)\,dx\,dz \\
&= \int_{-\infty}^{\infty} z\phi(z) \int_{0}^{z} g'(x)\,dx\,dz \\
&= \int_{-\infty}^{\infty} z\phi(z)(g(z) - g(0))\,dz \\
&= \int_{-\infty}^{\infty} z\phi(z)g(z)\,dz - \int_{-\infty}^{\infty} z\phi(z)g(0)\,dz \\
&= \int_{-\infty}^{\infty} z\phi(z)g(z)\,dz \\
&= \int_{-\infty}^{\infty} x\phi(x)g(x)\,dx \\
&= \mathbf{E}[X\,g(X)] \qquad (3.13)
\end{aligned}
$$

3.2.2 Arbitrary Normal Distribution: $Y \sim N(\mu, \sigma^2)$. Using the results of the last section, we will extend equation 3.13 to the case of an arbitrary normal random variable. The results of this section will be used in the general case of a normal random vector whose components are independent normal variables with the same variance and arbitrary mean.

Let Y be a real random variable with an arbitrary normal distribution. Since $Y \sim N(\mu, \sigma^2)$, the random variable $X = \frac{(Y-\mu)}{\sigma}$ has a standard normal distribution (i.e, $X \sim N(0,1)$). Define $h : \mathbf{R} \longrightarrow \mathbf{R}$ such that

$$h(Y) = g\left[\frac{(Y-\mu)}{\sigma}\right], \tag{3.14}$$

where g is defined by equation 3.5. We shall derive a formula for $\mathbf{E}\big[h'(Y)\big]$.

$$\begin{aligned}
\mathbf{E}\big[h'(Y)\big] &= \mathbf{E}\left[\frac{dg}{dY}\left[\frac{(Y-\mu)}{\sigma}\right]\right] \\
&= \mathbf{E}\left[\frac{1}{\sigma}g'\left[\frac{(Y-\mu)}{\sigma}\right]\right] \\
&= \frac{1}{\sigma}\mathbf{E}\left[g'\left[\frac{(Y-\mu)}{\sigma}\right]\right] \\
&= \frac{1}{\sigma}\mathbf{E}\big[g'[X]\big] \\
&= \frac{1}{\sigma}\mathbf{E}\Big[X\, g[X]\Big] \\
&= \frac{1}{\sigma}\mathbf{E}\left[\frac{(Y-\mu)}{\sigma}g\left[\frac{(Y-\mu)}{\sigma}\right]\right] \\
&= \mathbf{E}\left[\frac{(Y-\mu)}{\sigma^2}h(Y)\right].
\end{aligned} \tag{3.15}$$

3.2.3 Generalized Formulas For A Multivariate Normal Distribution. The formulas we derived for the single normal random variables can be generalized to the case of a normal random vector in which each element is an independent normal random variable with the same variance σ^2.

3.2.3.1 Multidimensional Definitions And Notations. Let $\vec{X} = (X_0, X_1, X_2, \ldots, X_{N-1})$ be a normal random vector in which each element X_i is an independent normal random variable such that for $i = 0, 1, 2, \ldots, N-1$

$$X_i \sim N(\mu_i, \sigma^2)$$

The mean of the vector $\vec{X}$ is defined as

$$\vec{\mu} = (\mu_0, \mu_1, \mu_2, \ldots, \mu_{N-1}). \tag{3.16}$$

The energy of the normal random vector $\vec{X}$ is defined as

$$\|\vec{X}\|^2 = \sum_{i=0}^{N-1} X_i^2. \tag{3.17}$$

A function $h : \mathbf{R}^N \longrightarrow \mathbf{R}$ is called almost differentiable if there exists a function $\nabla h : \mathbf{R}^N \longrightarrow \mathbf{R}^N$ such that, for all $\vec{z} \in \mathbf{R}^N$

$$h(\vec{x} + \vec{z}) - h(\vec{x}) = \int_0^1 \vec{z} \cdot \nabla h(\vec{x} + t\vec{z})\, dt, \tag{3.18}$$

for almost all $\vec{x} \in \mathbf{R}^N$. A function $g : \mathbf{R}^N \longrightarrow \mathbf{R}^N$ is called almost differentiable if all its coordinates are. The symbol ∇ is the vector differential operator of first partial derivatives with i^{th} coordinate

$$\nabla_i = \frac{\partial}{\partial x_i},$$

so that

$$\nabla_i h(\vec{x}) = \frac{\partial h(\vec{x})}{\partial x_i}, \tag{3.19}$$

$$\nabla h(\vec{x}) = \left(\frac{\partial h(\vec{x})}{\partial x_0}, \frac{\partial h(\vec{x})}{\partial x_1}, \ldots, \frac{\partial h(\vec{x})}{\partial x_{N-1}} \right). \tag{3.20}$$

3.2.3.2 Basic Formulas For An Arbitrary Normal Multidimensional Random Variable.

Let $\vec{X}$ be the multidimensional normal random variable defined in the previous section and h : $\mathbf{R}^N \longrightarrow \mathbf{R}$ an almost differentiable function such that

$$\mathbf{E}\Big[\|\nabla h(\vec{X})\|\Big] \;<\; \infty, \tag{3.21}$$

where

$$\mathbf{E}\Big[\nabla h(\vec{X})\Big] = \mathbf{E}\Big[\Big(\nabla_0 h(\vec{X}), \nabla_1 h(\vec{X}), \ldots, \nabla_{N-1} h(\vec{X})\Big)\Big].$$

By analogy to equation 3.15, we can write for the multidimensional case

$$\mathbf{E}\Big[\nabla h(\vec{X})\Big] \;=\; \mathbf{E}\Big[\frac{(\vec{X}-\vec{\mu})}{\sigma^2} h(\vec{X})\Big]. \tag{3.22}$$

Since each component X_i (for $i = 0, 1, 2, \ldots, N-1$) is an independent normal random variable and $X_i \sim N(\mu_i, \sigma^2)$, we can write

$$\mathbf{E}\Big[\frac{\partial h(\vec{X})}{\partial X_i}\Big] \;=\; \mathbf{E}\Big[\frac{(X_i-\mu_i)}{\sigma^2} h(\vec{X})\Big]. \tag{3.23}$$

3.2.4 A Closed Form Of Stein's Error Function. Given a multidimensional normal vector $\vec{X}$, composed of independent normal random variables $X_i \sim N(\mu_i, \sigma^2)$ for $i = 0, 1, 2, \ldots, N-1$, Stein defined an estimate $\hat{\vec{\mu}} = (\hat{\mu}_0, \hat{\mu}_1, \ldots, \hat{\mu}_{N-1})$ of the true mean $\vec{\mu} = (\mu_0, \mu_1, \ldots, \mu_{N-1})$ as follows

$$\hat{\vec{\mu}} \;=\; \vec{X} + g(\vec{X}), \tag{3.24}$$

where $g : \mathbf{R^N} \longrightarrow \mathbf{R^N}$ is an almost differentiable function with coordinates $g(\vec{X}) = \big(g_0(\vec{X}), g_1(\vec{X}), \ldots, g_{N-1}(\vec{X})\big)$ such that

$$g_i : \mathbf{R^N} \longrightarrow \mathbf{R},$$

and

$$\mathbf{E}_\mu\left[\sum_{i=0}^{N-1}|\nabla_i g_i(\vec{X})|\right] < \infty,$$

where the subscript μ indicates the dependence of the expectation operator on the mean.

For each normal random variable X_i, Charles Stein, defined an **unbiased estimate of the risk (SURE)** associated with estimating the true mean μ_i of the **single** independent normal random variable X_i as the expected squared error between the estimate $\hat{\mu}_i$ and the true mean μ_i as follows

$$\begin{aligned}\mathbf{E}_\mu\left[(\hat{\mu}_i - \mu_i)^2\right] &= \mathbf{E}_\mu\left[(X_i + g_i(\vec{X}) - \mu_i)^2\right] \\ &= \mathbf{E}_\mu\left[(X_i - \mu_i)^2 + 2g_i(\vec{X})(X_i - \mu_i) + g_i^2(\vec{X})\right] \\ &= \mathbf{E}_\mu\left[(X_i - \mu_i)^2\right] + \mathbf{E}_\mu\left[g_i^2(\vec{X})\right] + 2\mathbf{E}_\mu\left[g_i(\vec{X})(X_i - \mu_i)\right].\end{aligned} \tag{3.25}$$

Since

$$\mathbf{E}_\mu\left[(X_i - \mu_i)g_i(\vec{X})\right] = \sigma^2\mathbf{E}_\mu\left[\frac{\partial g_i(\vec{X})}{\partial X_i}\right], \tag{3.26}$$

equation 3.25 becomes

$$\mathbf{E}_\mu\left[(\hat{\mu}_i - \mu_i)^2\right] = \sigma^2 + \mathbf{E}_\mu\left[g_i^2(\vec{X})\right] + 2\sigma^2\mathbf{E}_\mu\left[\frac{\partial g_i(\vec{X})}{\partial X_i}\right]. \tag{3.27}$$

Using the above equation for a single random variable, Charles Stein defined an unbiased estimate of the risk associated with estimating the mean $\vec{\mu}$ of the vector $\vec{X}$ as follows

$$\mathbf{E}_\mu\left[\|\hat{\vec{\mu}} - \vec{\mu}\|^2\right] = \sum_{i=1}^{N}\mathbf{E}_\mu\left[(\hat{\mu}_i - \mu_i)^2\right]$$

$$= \sum_{i=1}^{N} \sigma^2 + \mathbf{E}_\mu\left[g_i^2(\vec{X})\right] + 2\sigma^2 \mathbf{E}_\mu\left[\frac{\partial g_i(\vec{X})}{\partial X_i}\right]$$

$$= N\sigma^2 + \mathbf{E}_\mu\left[\|g(\vec{X})\|^2\right] + 2\sigma^2 \mathbf{E}_\mu\left[\nabla g(\vec{X})\right]. \quad (3.28)$$

Ideally, we want to minimize the risk defined by equation 3.28 in order to get a more accurate estimate of the mean. Since this equation depends on the choice of the function g, many different choices, which satisfy the differentiability conditions stated above, are available. Since the basic estimation technique is based on adding a value to each element of the random vector $\vec{X}$, the next section introduces two different choices of the function g. These choices have a lot of practical applications and can be used to de-noise signals degraded by additive white Gaussian noise. In particular, the theory of Stein, proves that for white Gaussian noise with **zero** mean and a variance of σ^2, the mean estimate using Stein's criteria, is theoretically, **zero**. In other words, when we input zero mean white Gaussian noise signal to a Stein based mean estimator, we expect the output signal to be **zero**. This observation can be used to de-noise signals corrupted by additive white Gaussian noise with zero mean and a variance of σ^2.

3.3 Soft Thresholding Technique

Let $\vec{X}$ be a multidimensional normal random vector whose elements are independent normal random variables with the same variance σ^2 and let its mean be the vector $\vec{\mu} = (\mu_0, \mu_1, \ldots, \mu_{N-1})$. Define an estimate of the mean $\vec{\mu}$ by $\hat{\vec{\mu}} = (\hat{\mu}_0, \hat{\mu}_1, \ldots, \hat{\mu}_{N-1})$ such that (5) (6) (9) (8) (7)

$$\hat{\vec{\mu}} = \vec{X} + g(\vec{X}),$$

where $g(\vec{X}) = \left(g_0(\vec{X}), g_1(\vec{X}), \ldots, g_{N-1}(\vec{X})\right)$ is as defined in equation 3.24.

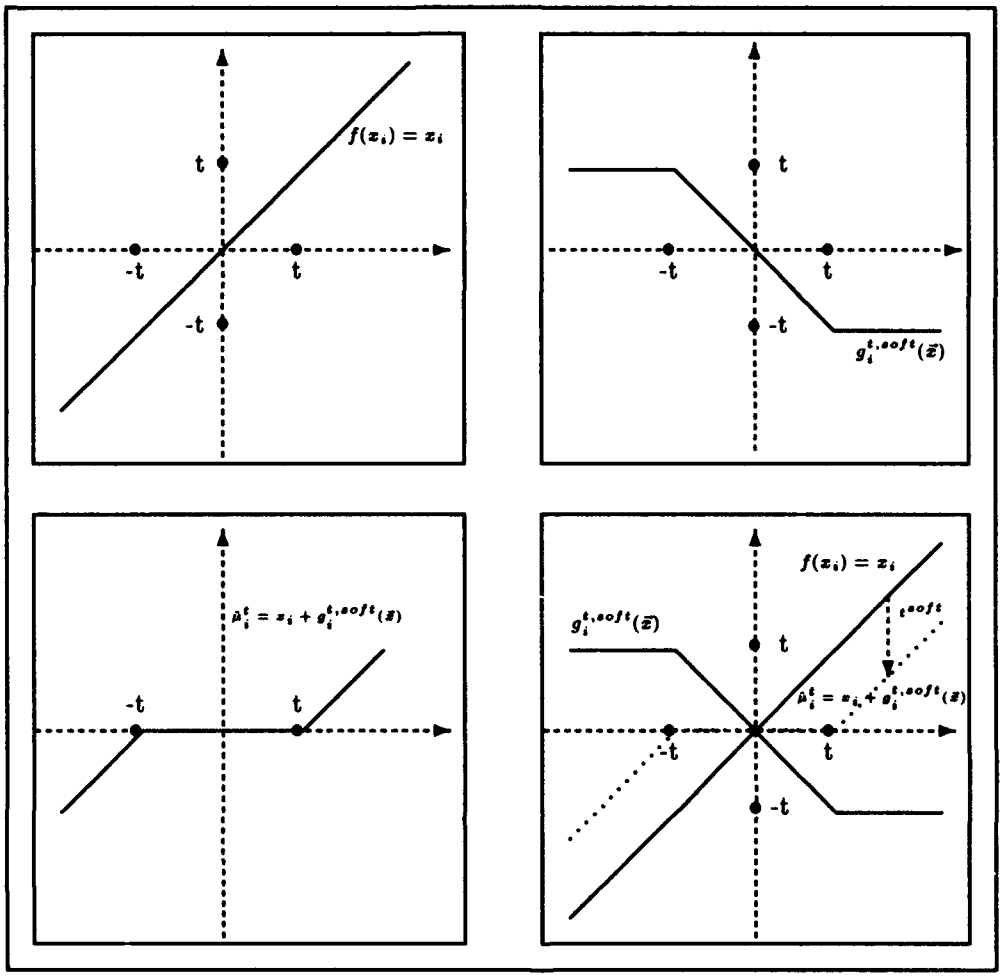

Figure 3.1 Soft thresholding technique (STT).

The Soft Thresholding Technique (STT) uses a threshold ($t \geq 0$) to estimate the true mean, μ_i, of each normal random variable, X_i, by the estimate $\hat{\mu}_i^t$, defined by (see figure 3.1)

$$\hat{\mu}_i^t = X_i + g_i^t(\vec{X}),$$

where for each $i = 0, 1, 2, \ldots, N-1$

$$g_i^t(\vec{X}) = \begin{cases} -t\,\mathrm{sgn}(X_i) & |X_i| > t \\ -X_i & |X_i| \leq t. \end{cases} \tag{3.29}$$

This yields

$$\hat{\mu}_i^t = \begin{cases} X_i - t\,\mathrm{sgn}(X_i) & |X_i| > t \\ 0 & |X_i| \le t. \end{cases} \tag{3.30}$$

An alternative representation of the soft thresholding technique is obtained by use of the minimum operator to write

$$g_i^{t,soft}(\vec{X}) = -\min\left(|X_i|, t\right)\mathrm{sgn}(X_i). \tag{3.31}$$

Then, for soft thresholding, the mean estimate is defined as

$$\hat{\mu}_i^t = X_i - \min\left(|X_i|, t\right)\mathrm{sgn}(X_i). \tag{3.32}$$

3.3.0.1 Definition of The Soft SURE Function. Since $g_i^t(\vec{X})$ is almost differentiable, we may write

$$\frac{\partial g_i^t(\vec{X})}{\partial X_i} = \begin{cases} 0 & |X_i| > t \\ -1 & |X_i| \le t. \end{cases} \tag{3.33}$$

By using the characteristic function which is defined by

$$\chi_{[-t,t]}(X_i) = \begin{cases} 0 & |X_i| > t \\ 1 & |X_i| \le t, \end{cases} \tag{3.34}$$

we get

$$\frac{\partial g_i^t(\vec{X})}{\partial X_i} = -\chi_{[-t,t]}(X_i). \tag{3.35}$$

We conclude then that

$$\nabla g^t(\vec{X}) = -\sum_{i=0}^{N-1} \chi_{[-t,t]}(X_i). \tag{3.36}$$

Since

$$\begin{aligned} \|g^t(\vec{X})\|^2 &= \sum_{i=0}^{N-1} [g_i^t(\vec{X})]^2 \\ &= \sum_{i=0}^{N-1} \Big[\min\big(|X_i|, t\big)\Big]^2, \end{aligned} \tag{3.37}$$

combining equations 3.24, 3.36, and 3.37 together, Donoho and Johnstone (5) (6) (9) obtained the following:

$$SURE_{soft}(t, \vec{X}) = \Big[N\sigma^2\Big] + \Big[\sum_{i=0}^{N-1} \Big[\min\big(|X_i|, t\big)\Big]^2\Big] - 2\sigma^2\Big[\sum_{i=0}^{N-1} \chi_{[-t,t]}(X_i)\Big]. \tag{3.38}$$

equation 3.28 becomes:

$$\mathbf{E}_\mu\Big[\|\hat{\vec{\mu}} - \vec{\mu}\|^2\Big] = \mathbf{E}_\mu\Big[SURE_{soft}(t, \vec{X})\Big]. \tag{3.39}$$

3.3.0.2 Soft Threshold. Since we want to minimize the estimate of the error associated with estimating the mean $\vec{\mu}$, we need to choose a threshold t^{soft} that minimizes the $SURE_{soft}$ quantity defined by equation 3.38. In order to choose the right threshold we need to proceed as follows. Assume that the coordinates X_i of the vector $\vec{X}$ have been ordered in an ascending manner by absolute value such that:

$$|X_0| \leq |X_1| \leq \ldots \leq |X_{N-1}|, \tag{3.40}$$

and let $t \geq 0$ be an arbitrary threshold such that for some $i = 0, 1, 2, \ldots, N-1$

$$|X_i| \leq t < t + \Delta t \leq |X_{i+1}|.$$

We have

$$\begin{aligned}
SURE_{soft}(t+\Delta t, \vec{X}) - SURE_{soft}(t, \vec{X}) &= \sum_{j=0}^{N-1} \left[\left[\min\left(|X_j|, t+\Delta t\right) \right]^2 - \left[\min\left(|X_j|, t\right) \right]^2 \right] - \\
&\quad 2\sigma^2 \left[\sum_{j=0}^{N-1} \chi_{[-(t+\Delta t), t+\Delta t]}(X_j) - \sum_{j=0}^{N-1} \chi_{[-t,t]}(X_j) \right] \\
&= \sum_{j=0}^{N-1} \left[\left[\min\left(|X_j|, t+\Delta t\right) \right]^2 - \left[\min\left(|X_j|, t\right) \right]^2 \right] \\
&= \sum_{j=0}^{i} \left[\left[\min\left(|X_j|, t+\Delta t\right) \right]^2 - \left[\min\left(|X_j|, t\right) \right]^2 \right] + \\
&\quad \sum_{j=i+1}^{N-1} \left[\left[\min\left(|X_j|, t+\Delta t\right) \right]^2 - \left[\min\left(|X_j|, t\right) \right]^2 \right] \\
&= \sum_{j=i+1}^{N-1} \left[\left[\min\left(|X_j|, t+\Delta t\right) \right]^2 - \left[\min\left(|X_j|, t\right) \right]^2 \right] \\
&= \sum_{j=i+1}^{N-1} \left[(t+\Delta t)^2 - t^2 \right] \\
&= \sum_{j=i+1}^{N-1} \left[(2t+\Delta t)\Delta t \right] \\
&> 0, \qquad (3.41)
\end{aligned}$$

which means that

$$SURE_{soft}(t+\Delta t, \vec{X}) > SURE_{soft}(t, \vec{X}) \geq SURE_{soft}(|X_i|, \vec{X}).$$

We conclude then that in order to choose a threshold that minimizes the $SURE_{soft}$ quantity, we need only test thresholds that are elements of the *known* set $\{|X_i|\}_{i=0}^{N-1}$.

The domain for our soft threshold is then defined as

$$t^{soft} \in \{0\} \cup \{|X_i|\}_{i=0}^{N-1}.$$

The value 0 is included in order to take care of the cases where $\hat{\vec{\mu}} = \vec{\mu}$ and $\sigma^2 = 0$. The threshold that minimizes the $SURE_{soft}$ quantity will be denoted by

$$t^{Soft} = \arg\Big[\min[SURE_{soft}(t, \vec{X})]\Big], \tag{3.42}$$

where $t \in \{0\} \cup \{|X_i|\}_{i=0}^{N-1}$.

3.4 Hard Thresholding Technique

Just like the Soft Thresholding Technique (STT), the Hard Thresholding Technique (HTT) uses a threshold ($t \geq 0$) to estimate the true mean, μ_i, of each independent normal random variable, X_i, by the estimate $\hat{\mu}_i^t$, defined by (see figure 3.2)

$$\hat{\mu}_i^t = X_i + g_i^t(\vec{X}),$$

where for each $i = 0, 1, 2, \ldots, N-1$

$$g_i^t(\vec{X}) = \begin{cases} 0 & |X_i| > t \\ -X_i & |X_i| \leq t. \end{cases} \tag{3.43}$$

This yields

$$\hat{\mu}_i^t = \begin{cases} X_i & |X_i| > t \\ 0 & |X_i| \leq t, \end{cases} \tag{3.44}$$

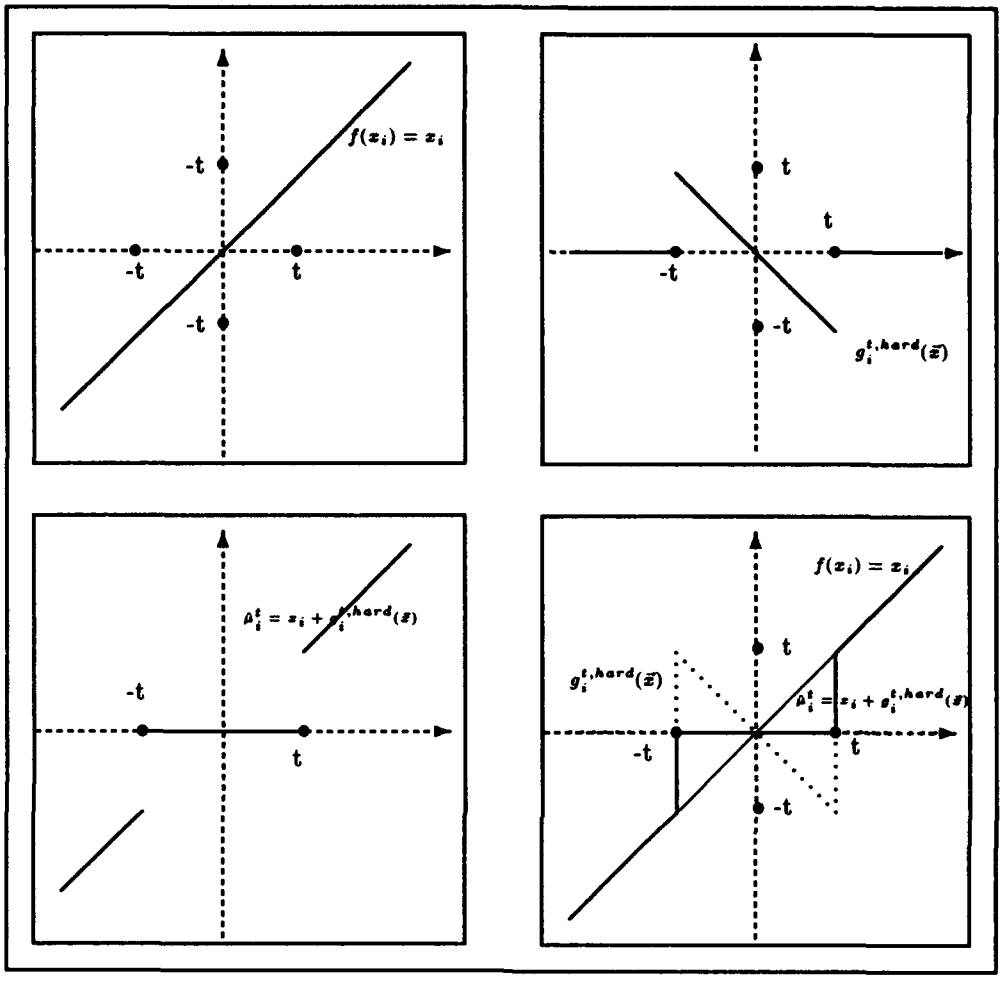

Figure 3.2 Hard thresholding technique (HTT).

An alternative representation of the hard thresholding technique (HTT) is obtained by use of the characteristic function defined by equation 3.34, such that

$$\hat{\mu}_i^t = X_i \chi_{[-t,t]}(X_i). \tag{3.45}$$

Then, for hard thresholding, the g_i^t function, is defined as

$$g_i^{t,hard}(\vec{X}) = X_i(1 - \chi_{[-t,t]}(X_i)). \tag{3.46}$$

3.4.0.3 Definition of The Hard SURE Function. Although the hard thresholding function, $g_i^t(\vec{X})$, is not almost differentiable, we decided to use it with Stein's criteria in order to compare the results with the soft thresholding technique. We may then write

$$\frac{\partial g_i^t(\vec{X})}{\partial X_i} = \begin{cases} 0 & |X_i| > t \\ -1 & |X_i| \leq t, \end{cases} \tag{3.47}$$

by using the characteristic function, we have

$$\frac{\partial g_i^t(\vec{X})}{\partial X_i} = -\chi_{[-t,t]}(X_i). \tag{3.48}$$

We conclude then that

$$\nabla g^t(\vec{X}) = -\sum_{i=0}^{N-1} \chi_{[-t,t]}(X_i). \tag{3.49}$$

Since

$$\begin{aligned} \|g^t(\vec{X})\|^2 &= \sum_{i=0}^{N-1} \Big[g_i^t(\vec{X}) \Big]^2 \\ &= \sum_{i=0}^{N-1} \Big[X_i^2 \cdot \chi_{[-t,t]}(X_i) \Big], \end{aligned} \tag{3.50}$$

combining equations 3.28, 3.49, and 3.50 together, we can define the following quantity

$$SURE_{hard}(t, \vec{X}) = \Big[N\sigma^2 \Big] + \Big[\sum_{i=0}^{N-1} \big[X_i^2 \cdot \chi_{[-t,t]}(X_i) \big] \Big] - 2\sigma^2 \Big[\sum_{i=0}^{N-1} \chi_{[-t,t]}(X_i) \Big], \tag{3.51}$$

equation 3.28 becomes:

$$\mathbf{E}_\mu \Big[\|\hat{\vec{\mu}} - \vec{\mu}\|^2 \Big] = \mathbf{E}_\mu \Big[SURE_{hard}(t, \vec{X}) \Big]. \tag{3.52}$$

Just like the case of the soft threshold, the domain of the hard threshold is given by:

$$t^{hard} \in \{0\} \cup \{|X_i|\}_{i=0}^{N-1}.$$

and the hard threshold should be chosen such that the $SURE_{hard}$ is minimized

$$t^{Hard} = \arg\left[\min[SURE_{hard}(t, \vec{X})]\right], \tag{3.53}$$

where $t \in \{0\} \cup \{|X_i|\}_{i=0}^{N-1}$.

3.5 Wavelet Transform

The continuous wavelet transform (CWT) is a technique that decomposes and analyzes a finite energy signal, $f(t) \in L^2(\mathbf{R})$ (set of Lebesgue-measurable functions) , using different resolutions for different scales (4), where

$$L^2(\mathbf{R}) = \left\{f \,|\, \int_{-\infty}^{+\infty} |f(t)|^2 \, dt < \infty\right\}. \tag{3.54}$$

The CWT is based on defining a "mother wavelet", ψ, which is subject to the following condition of admissibility:

$$\int_{-\infty}^{+\infty} |\xi|^{-1} |\tilde{\psi}(\xi)|^2 d\xi < \infty, \tag{3.55}$$

where $\tilde{\psi}$ is the Fourier transform of ψ. This condition implies that $\tilde{\psi}$ decays to zero as the frequency goes to infinity; furthermore, it implies that the mother wavelet, ψ, is zero-mean:

$$\int_{-\infty}^{+\infty} \psi(t) \, dt = 0. \tag{3.56}$$

Since equation 3.55 requires that the Fourier transform of $\psi(t)$ at the zero frequency (*i.e.*, $\omega = 0$) is zero

$$\tilde{\psi}(0) = 0, \tag{3.57}$$

it is clear that $\tilde{\psi}$ represents a band-pass filter (see figures B.1 through B.6 for three different wavelets).

Based on the above conditions, the continuous wavelet transform with scale a and shift b, is defined as

$$\mathcal{W}^{a,b}[f(t)] = \int_{-\infty}^{+\infty} f(t)\psi^*_{a,b}(t)\,dt, \tag{3.58}$$

where $(a, b) \in \mathbf{R}^+ \times \mathbf{R}$ and

$$\psi_{a,b}(t) = a^{-1/2}\psi\left(\frac{t-b}{a}\right), \tag{3.59}$$

and the asterisk indicates complex conjugation. The families of functions $\psi_{a,b}$ define a basis for the family of finite energy functions $L^2(\mathbf{R})$.

3.5.1 Properties of The Wavelet Transform. The following properties of the wavelet transform will prove very useful in our future derivations of the discrete wavelet transform (DWT) and the extension of the thresholding techniques to the wavelet domain.

Linearity: $\forall\, \alpha, \beta \in \mathbf{R}$

$$\mathcal{W}^{a,b}[\alpha f(t) + \beta g(t)] = \alpha\mathcal{W}^{a,b}[f(t)] + \beta\mathcal{W}^{a,b}[g(t)] \tag{3.60}$$

Scaling: $\forall\, \lambda \in \mathbf{R} - \{0\}$

$$\mathcal{W}^{a,b}\left[f\left(\frac{t}{\lambda}\right)\right] = \sqrt{|\lambda|}\mathcal{W}^{\frac{a}{\lambda},\frac{b}{\lambda}}[f(t)] \tag{3.61}$$

Time Shift: $\forall\, t_0 \in \mathbf{R}$

$$\mathcal{W}^{a,b}[f(t-t_0)] = \mathcal{W}^{a,b-t_0}[f(t)]. \tag{3.62}$$

From the above shifting property, it is clear that the wavelet transform is a linear time varying (LTV) operator.

3.5.2 Resolution Properties Of The Wavelet ψ. The energy of the Fourier transform of the wavelet ψ is defined as

$$\|\tilde{\psi}\|^2 = \int_{-\infty}^{+\infty} |\tilde{\psi}(f)|^2\, df. \tag{3.63}$$

If we normalize the wavelet ψ such that $\|\tilde{\psi}\|^2 = 1$, we have

$$\int_{-\infty}^{+\infty} |\tilde{\psi}(f)|^2\, df = 1, \tag{3.64}$$

which means that the square of the wavelet magnitude, $|\tilde{\psi}|^2$, represents a probability density function (pdf). Using the identities of Plancherel (preservation of energy) and Parserval (preservation of geometry), we can also define a time domain pdf as $|\psi(t)|^2$ such that

$$\int_{-\infty}^{+\infty} |\psi(t)|^2\, dt = 1. \tag{3.65}$$

We can then define the following statistics with respect to these wavelet based pdfs:

1. The **center frequency**, f_0, is defined as the expected value with respect to the pdf

$|\tilde{\psi}(f)|^2$, where f represents frequency.

$$f_0 = \int_0^{+\infty} f\,|\tilde{\psi}(f)|^2\,df. \tag{3.66}$$

2. **The second Moment** or the variance, σ_f^2, of this wavelet based pdf is then defined as

$$\sigma_f^2 = \int_0^{+\infty} (f - f_0)^2\,|\tilde{\psi}(f)|^2\,df. \tag{3.67}$$

The value of this variance measures the dispersion of frequencies relative to the mean f_0. The larger the variance, the more dispersed are the frequencies relative to the mean. This also means that the passband of the wavelet is larger with a wider bandwidth. The center frequency of a wavelet allows us to determine the range of frequencies that are filtered at a specific resolution level "a".

3.5.3 Resolution Properties Of The families of Wavelets $\psi_{a,b}$. The families of wavelets $\psi_{a,b}(t)$ are formed by dialations (using the scale a) and translations (using the shift b) of the mother wavelet ψ. The admissibility condition defined above still holds for these newly formed wavelets. Since

$$\psi_{a,b}(t) = a^{-1/2}\psi\Big(\frac{t-b}{a}\Big), \tag{3.68}$$

these wavelets have an expected value at time $t = b$ and it can be shown that their variance is given by

$$\sigma_{(a,b)}^2 = a^2\sigma_t^2, \tag{3.69}$$

where σ_t^2 is the variance of the mother wavelet.

1. **The Fourier transform** of $\psi_{a,b}(t)$ is given by

$$\tilde{\psi}_{a,b}(f) = \sqrt{a}e^{-i2\pi fb}\tilde{\psi}(af), \tag{3.70}$$

where $\tilde{\psi}$ is the Fourier transform of the mother wavelet, defined by

$$\tilde{\psi}(f) = \int_{-\infty}^{+\infty} \psi(t)e^{-i2\pi ft}\,dt, \tag{3.71}$$

and i is the complex number such that

$$i^2 = -1. \tag{3.72}$$

2. **The center frequency**, $f_{a,b}$, of these wavelets is related to the center frequency, f_0, of the mother wavelet by the following relation

$$f_{a,b} = \frac{f_0}{a}. \tag{3.73}$$

3. **The variance** of these wavelets, $\sigma_{a,b}^2$, are then related to the variance, σ_f^2, of the mother wavelet by the following equation

$$\sigma_{a,b}^2 = \frac{\sigma_f^2}{a^2}. \tag{3.74}$$

A moment's reflection on the above two parameters shows that as the value of the dilation parameter a increases, the bandpass center frequency ,$f_{a,b}$, of the wavelet $\tilde{\psi}_{a,b}(t)$, approaches the lower frequencies near the origin, the dc frequency, with a smaller variance or bandwidth, $\frac{\sigma_f^2}{a^2}$. This

shows that by changing the value of the dilation parameter "a", we can "zoom in" to different frequencies in the spectrum of the signal $f(t)$.

3.6 Discrete Wavelet Transform

Since the admissibility condition defined above holds for $\psi_{a,b}(t)$, the families of functions $\psi_{a,b}(t)$, which are formed by dialations (scale a) and translations (shift b) of the mother wavelet ψ, are themselves wavelets. They form a basis for $L^2(\mathbf{R})$. Since equation 3.58 represents an inner product between the function $f(t)$ and the corresponding wavelet $\psi_{a,b}(t)$, the wavelet transform with a particular choice of "a" and "b" is, indeed, a measure of the similarity between $f(t)$ and $\psi_{a,b}(t)$. While these newly formed wavelets are a basis for $L^2(\mathbf{R})$, they are not necessarily orthogonal and may redundantly represent the signal, $f(t)$ (1). By discretizing the values of the shift and scale parameters, it is possible to find an orthonormal set of wavelets to represent functions in $L^2(\mathbf{R})$. If we choose $a = a_0^m$ and $b = nb_0a_0^m$ for some $m, n \in \mathbf{Z}$, it is possible to find an orthonormal wavelet basis for $L^2(\mathbf{R})$. The choice most commonly made is for $a_0 = 2$ and $b_0 = 1$, where a_0 is known as the dilation factor.

3.6.1 Multi-resolution Analysis. The Multi-resolution Analysis (MRA) of a signal $f(t)$ was first introduced by Mallat and Meyer in 1986 (16). The MRA decomposes a signal into a set of approximations where the orthonormal wavelet bases are used as a tool to describe, mathematically, the "increment of information" needed to go from one coarse approximation to a finer or higher resolution approximation (4). Since the analysis of the signal $f(t)$ is based on a set of orthonormal wavelets which form a basis for $L^2(\mathbf{R})$, the amount of information needed to implement the MRA is kept to a minimum. Mallat developed a fast algorithm to implement the MRA.

3.6.1.1 MRA Requirements. A multi-resolution analysis consists of a set of approximation spaces, $V_j \subset L^2(\mathbf{R})$ ($j \in \mathbf{Z}$), which satisfy the following six requirements (21):

Requirement 1

The approximation spaces V_j are embedded such that

$$\ldots \subset V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \subset \ldots \tag{3.75}$$

Requirement 2

$$\overline{\bigcup_{j \in \mathbf{Z}} V_j} = L^2(\mathbf{R}). \tag{3.76}$$

Requirement 3

$$\bigcap_{j \in \mathbf{Z}} V_j = \{0\}. \tag{3.77}$$

Equation 3.76 ensures that $\forall\ f \in L^2(\mathbf{R})$

$$\lim_{j \to -\infty} P_j f = f,$$

where $P_j f$ is the orthogonal projection of $f(t)$ onto V_j.

Requirement 4

The above **approximation spaces** must satisfy

$$f(t) \in V_j \Longleftrightarrow f(2^j t) \in V_0. \tag{3.78}$$

Equations 3.75 and 3.78 imply that all spaces of the MRA are scaled versions of the central space V_0.

Requirement 5

The central space V_0 must be **invariant** under integer translations. $\forall n \in \mathbf{Z}$ we have

$$f(t) \in V_0 \Longrightarrow f(t-n) \in V_0. \tag{3.79}$$

Requirement 6

There must exist a *scaling* function $\phi \in V_0$ such that

$$\{\phi_{0,n}\}_{n\in\mathbf{Z}} \text{ is an orthonormal basis in } V_0,$$

where $\forall m, n \in \mathbf{Z}$

$$\phi_{m,n}(x) = 2^{-m/2}\phi(2^{-m}x - n). \tag{3.80}$$

The above equation implies that the set $\{\phi_{m,n}\}_{n\in\mathbf{Z}}$ is an orthonormal basis for the approximation space V_m, where $m \in \mathbf{Z}$.

3.6.1.2 Detail spaces. To completely characterize the MRA, the above six criteria can be used to construct a set of orthonormal wavelet basis $\{\psi_{m,n}\}_{n,m\in\mathbf{Z}}$ of $L^2(\mathbf{R})$, where

$$\psi_{m,n}(x) = 2^{-m/2}\psi(2^{-m}x - n), \tag{3.81}$$

such that

$$P_{m-1}f = P_m f + \sum_{n\in\mathbf{Z}} \langle f, \psi_{m,n}\rangle, \tag{3.82}$$

where $P_m f$ is the orthogonal projection of f onto the approximation space V_m and $\langle f, \psi_{m,n}\rangle$ represents the $L^2(\mathbf{R})$ inner product of f and $\psi_{m,n}$.

Let W_m be the orthogonal complement of V_m in V_{m-1} such that

$$W_m \perp V_m, \text{ with } V_m \subset V_{m-1} \text{ and } W_m \subset V_{m-1}.$$

The above definitions imply that the orthogonal projection of the function $f(t)$ onto the approximation space V_{m-1} is the same as the orthogonal projection of the function $f(t)$ onto the approximation space V_m plus the "information difference", $Q_m f$, between the two successive approximations, $P_m f$

and $P_{m-1}f$:

$$Q_m f = P_{m-1}f - P_m f, \tag{3.83}$$

where $Q_m f \in W_m$ and $Q_m f \perp V_m$.

Equation 3.83 implies that the set $\{\psi_{m,n}\}_{n\in\mathbf{Z}}$ is an orthonormal basis for W_m and that

$$V_{m-1} = V_m \oplus W_m, \tag{3.84}$$

where $\oplus$ designates the direct sum operator of two linear spaces. Furthermore, the orthogonal complements, $\{W_m\}_{m\in\mathbf{Z}}$ are mutually orthogonal such that for $i \neq j$

$$W_i \perp W_j = 0.$$

Since the subspaces $\{W_m\}_{m\in\mathbf{Z}}$ are mutually orthogonal, they effectively divide $L^2(\mathbf{R})$ into mutually orthogonal subspaces and we have

$$\bigoplus_{m\in\mathbf{Z}} W_m = L^2(\mathbf{R}). \tag{3.85}$$

In conclusion, the set of wavelets $\{\psi_{m,n}\}_{n,m\in\mathbf{Z}}$ is an orthonormal basis for $L^2(\mathbf{R})$.

3.6.2 Decomposition and Reconstruction of a finite energy signal using DWT. Let $f(t) \in L^2(\mathbf{R})$, and denote the orthogonal projection of $f(t)$ onto the space W_m by $Q_m f(t)$. Since $\{\psi_{m,n}\}_{n\in\mathbf{Z}}$ is an orthonormal basis for W_m, we can write $Q_m f(t)$ as a linear combination of the discrete wavelet series $\{\psi_{m,n}\}_{n\in\mathbf{Z}}$ such that

$$Q_m f(t) = \sum_{n\in\mathbf{Z}} d_{m,n}\psi_{m,n}(t), \tag{3.86}$$

where $d_{m,n} = \langle f, \psi_{m,n} \rangle$ are known as the m^{th}-level "detail coefficients". Since $\{\phi_{m,n}\}_{n \in \mathbf{Z}}$ is an orthonormal basis for V_m, the orthogonal projection $P_m f(t)$ of $f(t)$ onto the space V_m is defined in a similar way as

$$P_m f(t) = \sum_{n \in \mathbf{Z}} c_{m,n} \phi_{m,n}(t), \tag{3.87}$$

where $c_{m,n} = \langle f, \phi_{m,n} \rangle$ are known as the m^{th}-level "approximation coefficients".

Consider the scaling function $\phi_{1,0}(t)$. Since $V_1 \subset V_0$, we can represent $\phi_{1,0}(t)$ as a linear combination of the *zeroth* level basis, $\{\phi_{0,n}(t)\}_{n \in \mathbf{Z}}$

$$2^{-1/2} \phi(t/2) = \sum_{n \in \mathbf{Z}} h_n \phi(t - n), \tag{3.88}$$

where

$$h_n = \langle \phi_{1,0}, \phi_{0,n} \rangle. \tag{3.89}$$

Similarly, since $W_1 \subset V_0$ and $\{\psi_{1,n}(t)\}_{n \in \mathbf{Z}}$ is a basis for W_1, we can define

$$2^{-1/2} \psi(t/2) = \sum_{n \in \mathbf{Z}} g_n \phi(t - n), \tag{3.90}$$

where

$$g_n = \langle \psi_{1,0}, \phi_{0,n} \rangle. \tag{3.91}$$

The discrete filters h_n and g_n play a major role in the multi-resolution analysis. Mallat showed that the h and g filters can be used to relate the approximations at the m^{th}-level to the approximations and details at the $(m+1)^{st}$-level, respectively. Using these filters, it can be shown

that the equations that relate the approximations and details of different levels are given by

$$c_{m,n} = \sum_{k \in \mathbf{Z}} c_{m-1,k} h_{k-2n} \tag{3.92}$$

$$d_{m,n} = \sum_{k \in \mathbf{Z}} c_{m-1,k} g_{k-2n}. \tag{3.93}$$

The above equations are the heart of the MRA fast algorithm that was developed by Mallat. Using these equations, we can calculate the approximations of the m^{th}-level using both the approximations and details of the $(m+1)^{st}$-level, as follows.

After decomposing the approximation coefficients at the m^{th}-level into details and approximations at the $(m+1)^{st}$-level, we can perform the inverse procedure by using these $(m+1)^{st}$-level approximations and details to get back our m^{th}-level approximations. In fact, the filters h and g may also be used to calculate the approximations at the m^{th}-level starting with both the approximations and details of the $(m+1)^{st}$-level using the following equation

$$c_{m-1,n} = \sum_{k \in \mathbf{Z}} c_{m,k} h_{n-2k} + \sum_{k \in \mathbf{Z}} d_{m,k} g_{n-2k}. \tag{3.94}$$

3.6.3 Characteristics Of The h and g Filters. Daubechies (4) showed that the filters h and g have the following properties

$$\sum_{n \in \mathbf{Z}} |h_n| < \infty \tag{3.95}$$

$$\sum_{n \in \mathbf{Z}} |g_n| < \infty. \tag{3.96}$$

The above two equations require that the filters h and g must be stable.

Let $H(f)$ and $G(f)$ represent the Fourier transforms of the filters h and g, respectively. A sufficient

condition for the construction of the ψ is that the matrix

$$U = \begin{bmatrix} H(f) & G(f) \\ H(f+\frac{1}{2}) & G(f+\frac{1}{2}) \end{bmatrix}, \tag{3.97}$$

must be unitary (i.e., $\overline{U}^{\top} U = I$, where I is the identity operator).

One possible choice for G is

$$G(f) = e^{-i2\pi f}\overline{H\left(f+\frac{1}{2}\right)}, \tag{3.98}$$

which lead to the following relation between the coefficients of the h and g filters, $\forall n \in \mathbf{Z}$

$$g_n = (-1)^{(1-n)}\overline{h_{1-n}}. \tag{3.99}$$

Finally, The filters h and g must satisfy the following conditions

$$\sum_{n\in\mathbf{Z}} h_n = \sqrt{2} \tag{3.100}$$

$$\sum_{n\in\mathbf{Z}} h_n^2 = 1 \tag{3.101}$$

$$\sum_{n\in\mathbf{Z}} g_n = 0 \tag{3.102}$$

$$\sum_{n\in\mathbf{Z}} g_n^2 = 1. \tag{3.103}$$

Equation 3.100 implies that the h filter is a low-pass filter while equation 3.102 implies that the g filter is a high-pass filter.

3.6.4 Examples Of Wavelets And Filter Coefficients. The following wavelets will be used in our analysis of noisy speech data (chapter 4). In tables A.1 through A.3, we present the filter coefficients of three different wavelets, db6, coiflet(6), and db20. These wavelet-based discrete filters have different filtering properties (see figures A.1, A.2, and A.3). Observe that the h filters are

low-pass filters, while the g filters are high-pass filters. Figures B.1 through B.6 show the wavelets, scaling functions, and their Fourier transforms. Observe, the amplitude of the Fourier transform of all wavelets represent band-pass filters; while the corresponding scaling functions represent low-pass filters. Notice, the wavelets corresponding to db6 and coiflet(6) have many high energy side-lobes; while those of the db20 wavelet, have very small side-lobes.

3.7 Implementation Of The Discrete Wavelet Transform (DWT)

In order to efficiently implement The MRA developed by S. Mallat, we proceed as follows (21)

Given a T-periodic signal $f(t)$ such that $\forall\, t \in \mathbf{R}$

$$f(t+T) = f(t), \tag{3.104}$$

the wavelet transform satisfies

$$\mathcal{W}^{a,b}[f(t+T)] = \mathcal{W}^{a,b+T}[f(t+T)], \tag{3.105}$$

which means that the continuous wavelet transform of a T-periodic signal, is also T-periodic. We can use this property to minimize the number of calculations needed to decompose a given signal into sets of details and sets of approximations. The next two sections use this property to develop an efficient algorithm for decomposing and reconstructing a signal using wavelets.

3.7.1 Decomposition Using DWT. Now, given the filter sequence h_n and N samples of the function $f(t)$, at a sampling period, Δt, we compute the approximation coefficients, $\{c_{m,n}\}_{n\in\mathbf{Z}}$ where $1 \leq m \leq M$, for a total of M levels of decomposition as

$$c_{m,n} = \sum_{k\in\mathbf{Z}} c_{m-1,k} h_{k-2n}, \tag{3.106}$$

where the *zeroth*-level approximation coefficients are taken to be the samples of $f(t)$ at integer multiples of Δt

$$c_{0,n} = f(n\Delta t).$$

Using equation 3.99, we can calculate the g_k filter sequence. The detail coefficients are then calculated using the following equation

$$d_{m,n} = \sum_{k \in \mathbf{Z}} c_{m-1,k} g_{k-2n}. \tag{3.107}$$

We can then write

$$c_{m,n} = \sum_{k \in \mathbf{Z}} c_{m-1,k} \tilde{h}_{j-k}|_{j=2n} \tag{3.108}$$

$$d_{m,n} = \sum_{k \in \mathbf{Z}} c_{m-1,k} \tilde{g}_{j-k}|_{j=2n}, \tag{3.109}$$

where $\forall\ n \in \mathbf{Z}$, the new filters $\tilde{h}$ and $\tilde{g}$ are defined as

$$\tilde{h}_n = h_{-n} \text{ and } \tilde{g}_n = g_{-n}.$$

The above two decomposition equations may be viewed as a two steps operation: A convolution of the sequence $\{c_{m-1,n}\}_{n \in \mathbf{Z}}$ with the filters $\tilde{h}$ and $\tilde{g}$, followed by the operation of "down-sampling" by a factor of 2; i.e., the convolutions are evaluated at $2n$, keeping only the evenly-indexed coefficients of the convolution's result.

If the filter h has at most L non-zero elements and the sampled signal $f_n = c_{0,n}$ has at most N non-zero elements, for $n = 0, 1, ..., N-1$, it can be shown that the above convolutions of h and g with $c_{0,n}$, will have, in general, $N + L - 1$ non-zero elements. The above convolution operations "spread" the sequences $c_{m,n}$ and $d_{m,n}$. In fact the spreading increases as we move from the m^{th} to the $(m+1)^{st}$-level, for $m = 1, 2, ..M$. In order to avoid this "spreading" at each stage of the decomposition, the *DWT* can be implemented using a periodic extension of f so that the sequence

$c_{0,n}$ is N-periodic

$$c_{0,n+N} = c_{0,n}.$$

Assuming that $N = 2^M$, where M is a positive integer, and due to the down-sampling operations mentioned above, the sequences $c_{m,n}$ and $d_{m,n}$ are also periodic with period $2^{-m}N$. We can then write the following relations for $m = 1, 2, \ldots, M$

$$c_{m,n} = c_{m,[n+2^{-m}N]}$$

$$d_{m,n} = d_{m,[n+2^{-m}N]}.$$

Starting with $N = 2^M$ samples of the original N-periodic signal, the down-sampled discrete wavelet transform (DWT) allows a maximum of M levels of decomposition where at each level m, we have exactly $2^{-m}N$ unique approximation coefficients ($c_{m,n}$) and $2^{-m}N$ unique detail coefficients ($d_{m,n}$). The last level of decomposition, the M^{th} level or the coarsest level, has one approximation element and one detail element (i.e., $2^{-M}N = 1$). After M levels of decomposition, we end-up with a total of $N - 1$ unique approximation coefficients and $N - 1$ unique detail coefficients (see figure 3.3).

To completely define the above convolutions, at each level m, we need only compute the $2^{-m}N$ unique elements. In order to efficiently implement the above convolutions, we can rewrite the approximation coefficients at the m^{th} decomposition level as

$$c_{m,n} = \sum_{k=k_s}^{k_e} c_{m-1,[(k+2n) \bmod (2^{-m}N)]} h_k,$$

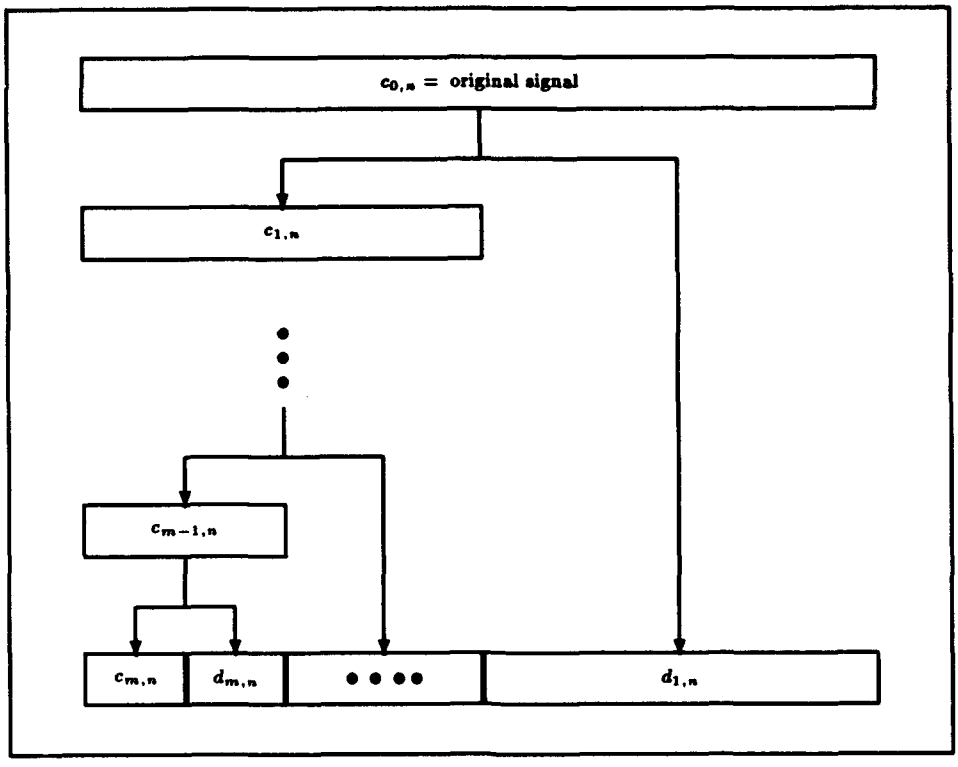

Figure 3.3 Wavelet decomposition of a signal staring with $N = 2^M$ samples and decomposing up to the m^{th}-level where $1 \leq m \leq M$.

where mod represent the modulo operator and k_s and k_e represent the first and last non-zero components of the filter h, respectively. They are related to the length, L of the filter h as follows

$$k_e - k_s = L - 1.$$

In a similar fashion, the detail coefficients are implemented as

$$d_{m,n} = \sum_{k=k_s}^{k_e} c_{m-1,[(k+2n) \bmod (2^{-m}N)]} g_k.$$

Since $g_n = (-1)^{(1-n)}\overline{h_{1-n}}$, the g filter length is also L and we have

$$\acute{k}_e - \acute{k}_s = L - 1.$$

The first and last non-zero elements of the filters h and g, can be chosen so that the filters' energies are well-centered, though not all wavelets have filters which can be centered exactly.

3.7.2 Reconstruction Using DWT. We have seen that the reconstruction of the approximation coefficients at the $(m-1)^{st}$-level are related to both the approximations and details of the m^{th}-level by

$$c_{m-1,n} = \sum_{k \in \mathbf{Z}} c_{m,k} h_{n-2k} + \sum_{k \in \mathbf{Z}} d_{m,k} g_{n-2k}, \tag{3.110}$$

where for M levels of decompositions, m takes the values $m = 1, 2, \ldots M$.

The above equation can be rewritten as

$$c_{m-1,n} = \sum_{k \in \mathbf{Z}} \tilde{c}_{m,k} h_{n-k} + \sum_{k \in \mathbf{Z}} \tilde{d}_{m,k} g_{n-k},$$

wherein $\tilde{c}_{m,k}$ and $\tilde{d}_{m,k}$ represent the "up-sampled" approximation and detail coefficients at the m^{th} decomposition level, respectively. $\forall\ k \in \mathbf{Z}$

$$\tilde{c}_{m,2k} = c_{m,k} \text{ and } \tilde{c}_{m,2k+1} = 0$$

$$\tilde{d}_{m,2k} = d_{m,k} \text{ and } \tilde{d}_{m,2k+1} = 0.$$

In order to efficiently implement the above reconstruction equation, using the periodic extension from the last section, we proceed as follows:

Since we have one unique approximation and one unique detail elements at the M^{th}-level (i.e., The M^{th}-level is 1-periodic, we can use the above equation to compute the approximations at the level above (i.e., $(M-1)^{st}$-level). The number of unique approximation coefficients is $2^{(M-1)}N$, where $N = 2^M$ is the number of samples we started with. We can then compute the approximations at the $(M-2)^{nd}$-level using this newly reconstructed approximation set and the $2^{(M-1)}N$ details obtained during the decomposition process at the $(M-1)^{st}$-level. All in all, for perfect reconstruction of

the sequence $\{c_{0,n}\}_{0 \leq n < N = 2^M}$ at the *zeroth* level, we need to keep the following data

1. All the details obtained during the decomposition process (a total of $N - 1$ unique detail coefficients).

2. The unique approximation coefficient obtained during the decomposition process at the M^{th} decomposition level.

In conclusion, starting with an $N = 2^M$-periodic signal, the full DWT (i.e., M levels of decomposition), produces a total of $N - 1$ unique detail coefficients, and 1 unique approximation coefficient at the M^{th} decomposition level, for a total of N coefficients. The partial DWT (i.e., m levels of decomposition where $1 \leq m \leq M$), produces a total of $N - 2^{(M-m)}$ unique detail coefficients, and $2^{(M-m)}$ unique approximation coefficient at the m^{th} decomposition level, for a total of $N = 2^M$ coefficients (see figure 3.4).

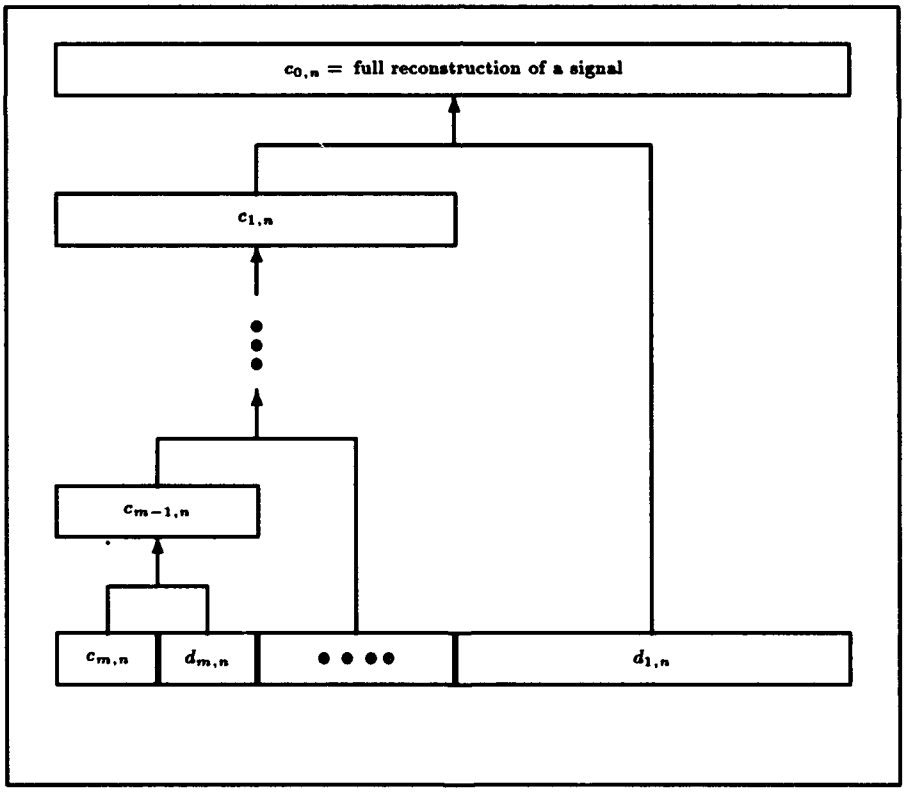

Figure 3.4 Wavelet reconstruction starting from the m^{th}-level where $1 \leq m \leq M$ to the *zeroth* level where the number of samples is $N = 2^M$

The reconstruction equation at the m^{th}-level can be rewritten as

$$c_{m-1,n} = \sum_{k=k_s}^{k_e} \left[(n-k+1) \bmod 2\right] c_{m,\left[\left[\frac{n-k}{2}\right] \bmod [2^{-m}N]\right]} h_k + \sum_{k=k_s}^{k_e} \left[(n-k+1) \bmod 2\right] d_{m,\left[\left[\frac{n-k}{2}\right] \bmod [2^{-m}N]\right]} g_k,$$

where $0 \leq n < 2^{-(m-1)}N$.

3.7.3 Statistical Properties Of The Wavelet Coefficients Of Random Variables. Let $\vec{X} = (X_0, X_1, X_2, \ldots, X_{N-1})$ be a normal random vector of $N = 2^M$ independent random variables such that for $i = 0, 1, 2, \ldots, N-1$

$$X_i \sim N(\mu_i, \sigma^2), \tag{3.111}$$

where $\vec{\mu} = (\mu_0, \mu_1, \mu_2, \ldots, \mu_{N-1})$ is the vector mean of the normal random vector $\vec{X}$. We will show that, at each level of decomposition, the details and approximations are also normal random vectors such that the discrete wavelet decomposition at the m^{th}-level ($1 \leq m \leq M$) is given as in equations 3.106 and 3.107 by

$$C_{m,n} = \sum_{k \in \mathbf{Z}} C_{m-1,k} h_{k-2n} \tag{3.112}$$

$$D_{m,n} = \sum_{k \in \mathbf{Z}} C_{m-1,k} g_{k-2n}, \tag{3.113}$$

where C and D denote the approximation and detail random variables, respectively. This property of the DWT coefficients allows us to use the SURE criteria which requires the input data to be normally distributed (see figures C.1 through C.6 for using the STT technique with a noisy sinewave).

During the decomposition process, the *zeroth* level approximations are taken to be the vector $\vec{X}$ itself such that

$$C_{0,n} = X_n,$$

where $n = 0, 1, 2, \ldots, N-1$. Since, according to equation 3.111, the vector $\vec{X}$ is normal and all random variables, X_i , are independent, the *zeroth*-level approximations are also independent and normally distributed with the same parameters as the vector $\vec{X}$. By using equation 3.112, the *first* level approximations can be written as a linear combination of the *zeroth* level approximations such that

$$C_{1,n} = \sum_{k \in \mathbf{Z}} C_{0,k} h_{k-2n}. \tag{3.114}$$

Since $C_{0,k} \sim N(\mu^C_{0,k}, \sigma^2)$, where $\mu^C_{0,k} = \mu_k$, we conclude that $C_{1,n}$ is also independent and normally distributed. The mean of $C_{1,n}$, denoted by $\mu^C_{1,n}$, is given by

$$\begin{aligned} \mu^C_{1,n} &= \mathbf{E}[C_{1,n}] \\ &= \mathbf{E}\left[\sum_{k \in \mathbf{Z}} C_{0,k} h_{k-2n}\right] \\ &= \sum_{k \in \mathbf{Z}} h_{k-2n} \mathbf{E}[C_{0,k}] \\ &= \sum_{k \in \mathbf{Z}} h_{k-2n} \mu^C_{0,k} \\ &= \sum_{k \in \mathbf{Z}} h_{k-2n} \mu_k. \end{aligned} \tag{3.115}$$

Using equation 3.101 and the independence of the *zeroth*-level approximations, $C_{0,n}$, the variance is given by (12)

$$\begin{aligned} \mathbf{Var}[C_{1,n}] &= \mathbf{Var}\left[\sum_{k \in \mathbf{Z}} C_{0,k} h_{k-2n}\right] \\ &= \sum_{k \in \mathbf{Z}} h^2_{k-2n} \mathbf{Var}[C_{0,k}] \end{aligned}$$

$$
\begin{aligned}
&= \sum_{k\in\mathbf{Z}} h_{k-2n}^2 \sigma^2 \\
&= \sigma^2 \sum_{k\in\mathbf{Z}} h_{k-2n}^2 \\
&= \sigma^2. \qquad (3.116)
\end{aligned}
$$

The random variable $C_{1,n}$ is then distributed as

$$C_{1,n} \sim N\Big(\sum_{k\in\mathbf{Z}} h_{k-2n}\mu_{0,k}^{C}\,,\ \sigma^2\Big). \qquad (3.117)$$

Recursively, the approximation coefficients at the m^{th}-level are also independent and normally distributed such that the mean is related to the mean of the $(m-1)^{st}$-level by

$$\mu_{m,n}^{C} = \sum_{k\in\mathbf{Z}} h_{k-2n}\mu_{m-1,k}^{C}. \qquad (3.118)$$

Using the above results, we can write

$$C_{m,n} \sim N(\mu_{m,n}^{C}\,,\ \sigma^2), \qquad (3.119)$$

where $\mu_{m,n}^{C}$ is defined by equation 3.118.

Since the detail coefficients are also a linear combination of the approximation coefficients (see equation 3.113), it is easy to show that the details at the m^{th}-level are also independent and normal random variables such that

$$D_{m,n} \sim N(\mu_{m,n}^{D}\,,\ \sigma^2), \qquad (3.120)$$

where the detail's means at the m^{th}-level are related to the approximation's means at the $(m-1)^{st}$-level by

$$\mu^D_{m,n} = \sum_{k \in \mathbf{Z}} g_{k-2n} \mu^C_{m-1,k}. \tag{3.121}$$

3.8 Complex Statistics and Analysis

The purpose of this section is to relate the statistics of a complex random variable to the statistics of its real and imaginary parts. The relations to be developed in this section, will be used in the analysis of the Fourier transform of normal random vector.

A complex number z can be defined in its rectangular form as

$$z = x + iy, \tag{3.122}$$

where x and y are real numbers which represent the real and imaginary parts of z, respectively. The complex number i is as defined in equation 3.72. The next sections, will develop the

3.8.0.1 Geometric Properties of Complex Numbers. The amplitude of a complex number is defined as

$$|z| = \sqrt{x^2 + y^2}. \tag{3.123}$$

When the product $xy \neq 0$, the phase of a complex number is defined as

$$\arg[z] = \theta, \tag{3.124}$$

where $0 \leq \theta < 2\pi$ and

$$\theta = \arctan\left(\frac{y}{x}\right). \tag{3.125}$$

a. If $z = 0$ (i.e., $x = 0$ and $y = 0$), then the phase is not defined.

b. If $z = x$ is pure real and non-zero (i.e., $x \neq 0$ and $y = 0$), then

$$\theta = \begin{cases} 0 & x > 0 \\ \pi & x < 0. \end{cases} \tag{3.126}$$

c. If $z = iy$ is imaginary and non-zero (i.e., $x = 0$ and $y \neq 0$), then

$$\theta = \begin{cases} \frac{\pi}{2} & y > 0 \\ \frac{3\pi}{4} & y < 0. \end{cases} \tag{3.127}$$

Using the above properties of complex number we can rewrite equation 3.122 in its polar form as

$$z = |z|e^{i\theta}, \tag{3.128}$$

where θ and $e^{i\theta}$ are defined by equations 3.125 and 3.140, respectively.

3.8.0.2 Statistical Properties Of Complex Random Variables. A complex random variable is defined as

$$Z = X + iY, \tag{3.129}$$

where both X and Y are real random variables.

1. The expected value of a complex random variable is defined as

$$\mathbf{E}[Z] = \mathbf{E}[X + iY]$$

$$= \mathbf{E}[X] + i\,\mathbf{E}[Y]. \tag{3.130}$$

2. The variance of a complex random variable is defined as:

$$\begin{aligned}
\mathbf{Var}[Z] &= \mathbf{Var}[X + iY] \\
&= \mathbf{E}[|Z|^2] - |\mathbf{E}[Z]|^2 \\
&= \mathbf{E}[X^2 + Y^2] - \left[\mathbf{E}[X]^2 + \mathbf{E}[Y]^2\right] \\
&= \left[\mathbf{E}[X^2] - \mathbf{E}[X]\right] + \left[\mathbf{E}[Y^2] - \mathbf{E}[X]\right] \\
&= \mathbf{Var}[X] + \mathbf{Var}[Y].
\end{aligned} \tag{3.131}$$

3.8.0.3 Statistics Of The Amplitude And Phase Of A Complex Random Variable.

Let Z be a complex random variable such that

$$Z = X + iY, \tag{3.132}$$

where both X and Y are real independent normal random variables

$$X \sim N(\mu_x, \sigma^2)$$

$$Y \sim N(\mu_y, \sigma^2).$$

The amplitude $|Z| = \sqrt{X^2 + Y^2}$, which is a function of the random variables X and Y, has a probability density function defined by

$$f(z) = \begin{cases} \frac{z}{\sigma^2} I_0\left[\frac{z\mu_x}{\sigma^2}\right] e^{\frac{-(z^2+\mu_x^2)}{2\sigma^2}} & z > 0 \\ 0 & z \leq 0, \end{cases} \tag{3.133}$$

where $\mu_z^2 = \mu_x^2 + \mu_y^2$ and $I_0(x)$ is the modified Bessel function defined as

$$I_0(x) = \frac{1}{2\pi} \int_0^{2\pi} e^{x \cos \alpha} d\alpha. \tag{3.134}$$

If $\mu_x = \mu_y = 0$, $f(z)$ is called a ***Rayleigh*** distribution (18). The phase θ of the complex random variable Z which is defined as

$$\theta = \arctan\left[\frac{Y}{X}\right], \tag{3.135}$$

where $-\pi < \theta \leq \pi$, has a uniform distribution (18) in the interval $(-\pi, \pi)$ defined by

$$f_\theta(\theta) = \begin{cases} \frac{1}{2\pi} & -\pi < \theta \leq \pi \\ 0 & otherwise. \end{cases} \tag{3.136}$$

3.9 Fourier Analysis

The purpose of this section is to define the discrete Fourier transform (DFT), apply the results of the last section to the real and imaginary parts of the DFT of a random vector, and define the statistics of the real and imaginary parts. The results of this section, will be used with the results of Stein in order to de-noise the real and imaginary parts of the DFT of a normal random vector.

Given a signal $f(t)$, one is interested in analyzing its frequency content ***locally in time*** (4). The standard Fourier transform which is defined as

$$(\mathcal{F}f)(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt, \tag{3.137}$$

gives a representation of the frequency content of $f(t)$, but it is unable to localize frequencies in time. In order to localize the time occurrence of many high frequency bursts, we may first window

the signal $f(t)$ and then take the Fourier transform of this windowed portion of the signal $f(t)$

$$(\mathcal{F}^{win} f)(\omega, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(s) g(s-t) e^{-i\omega s} \, ds, \tag{3.138}$$

where $g(t)$ is a window function.

The above equation is well known in the signal processing field by its discrete form, where the shift t and the frequency ω are discretized as $t = nt_0$ and $\omega = m\omega_0$. The Windowed Fourier transform or the short-time Fourier transform (STFT), $(\mathcal{F}^{win} f)(\omega, t)$, can be *interpreted* as the "amount of the frequency ω" present in the signal f near time t.

One similarity between the Fourier transform and the wavelet transform is that both equation 3.58 and 3.137 take the inner product of f with a family of functions indexed by two variables, $\psi_{a,b}(t) = a^{-1/2}\psi(\frac{t-b}{a})$ and $g^{\omega,t} = e^{i\omega s} g(s-t)$. However, the difference between the wavelet and windowed Fourier transforms lies in the shapes of the analyzing functions $g^{\omega,t}$ and $\psi_{a,b}(t)$.

The functions $g^{\omega,t}$ all consist of the same envelope function g, translated to the proper time location, and "filled in" with higher frequency oscillations. The windowed Fourier transform effectively divides the frequency spectrum of the function $f(t)$ into equal-bandwidth regions. In contrast, the windows used by the wavelet transform are well adapted to their frequency. The use of both a dilation factor "a" coupled with a shift variable "b", allows the wavelet transform to decompose and analyze signals using a small bandwidth (broader window) for low frequencies and large bandwidth (narrow window) for higher frequencies.

The main characteristic of the wavelet transform lies in its ability to "zoom in" and detect very short-lived high frequency phenomena, such as transients in signals or discontinuities in functions (i.e., human vocal tract glottal closure).

3.9.1 Discrete Fourier Transform (DFT). The discrete Fourier transform (DFT) of a periodic finite-length sequence of N points, $\{x_m\}_{m=0}^{N-1}$, is defined as

$$\tilde{x}_k = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} \left[x_m e^{i(\frac{2\pi}{N})km} \right], \tag{3.139}$$

where $0 \leq k \leq N-1$.

The quantity $e^{i\theta}$ is defined as

$$e^{i\theta} = \cos(\theta) + i \sin(\theta), \tag{3.140}$$

where i is the complex number defined by equation 3.72. For each $0 \leq k \leq N-1$, the quantity $|\tilde{x}_k|$ measures the amount of frequency $\omega = (\frac{2\pi}{N})k$ present in the signal $\{x_m\}_{m=0}^{N-1}$. In order to get back our original signal, $\{x_m\}_{m=0}^{N-1}$, from its DFT sequence, $\{\tilde{x}_k\}_{k=0}^{N-1}$, we perform the inverse discrete Fourier transform (IDFT) defined as

$$x_m = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \left[\tilde{x}_k e^{-i(\frac{2\pi}{N})km} \right]. \tag{3.141}$$

We conclude then that the sequence $\{x_m\}_{\{m=0,1,2,\ldots,N-1\}}$ can be represented as a sum of sinusoids of frequencies $0, 1, 2, \ldots, N-1$. Hence the discrete Fourier transform can also be interpreted as a frequency analysis (or *"spectrum analysis"*) of the input signal $\{x_m\}_{m=0}^{N-1}$ (19).

3.9.2 Properties Of The DFT. In this section, we will show some of the properties of the real and imaginary parts of the DFT of signal. We will use these properties in several occasions in order to decrease the number of calculations needed to implement the DFT. We will also show that some of the DFT components (i.e., dc component) have very unique properties.

Using equation 3.139 and 3.140, we can decompose the DFT into a *sine* and *cosine* series as follows

$$\begin{aligned}
\tilde{x}_k &= \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} \left[x_m e^{i(\frac{2\pi}{N})km} \right] \\
&= \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} \left[x_m \left[\cos(\frac{2\pi}{N}km) + i \sin(\frac{2\pi}{N}km) \right] \right] \\
&= \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} \left[x_m \cos\left[(\frac{2\pi}{N})km \right] \right] + \\
&\quad i \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} \left[x_m \sin\left[(\frac{2\pi}{N})km \right] \right].
\end{aligned} \tag{3.142}$$

From the above equation, the real part is defined as

$$\mathbf{Re}[\tilde{x}_k] = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} \left[x_m \cos\left[(\frac{2\pi}{N})km \right] \right], \tag{3.143}$$

and the imaginary part is defined as

$$\mathbf{Im}[\tilde{x}_k] = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} \left[x_m \sin\left[(\frac{2\pi}{N})km \right] \right]. \tag{3.144}$$

The elements of the DFT sequence, $\{\tilde{x}_k\}_{k=0}^{N-1}$, can then be rewritten as

$$\tilde{x}_k = \mathbf{Re}[\tilde{x}_k] + i\,\mathbf{Im}[\tilde{x}_k]. \tag{3.145}$$

Assume that N is even

a. The dc component ($k = 0$) is real:

$$\begin{aligned}
\tilde{x}_0 &= \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} [x_m] \\
&= \mathbf{Re}[\tilde{x}_0],
\end{aligned} \tag{3.146}$$

which means that the imaginary part of $\tilde{x}_0$ is zero:

$$\mathbf{Im}[\tilde{x}_0] = 0.$$

b. $x_{\frac{N}{2}}$ is real:

$$\tilde{x}_{\frac{N}{2}} = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} [(-1)^m x_m], \tag{3.147}$$

which means that the imaginary part of $\tilde{x}_{\frac{N}{2}}$ is also zero:

$$\mathbf{Im}[\tilde{x}_{\frac{N}{2}}] = 0.$$

c. Symmetry: $1 \leq k \leq N-1$

$$x_k = x_{N-k}. \tag{3.148}$$

The above equation has some practical consequences:

1. We need only calculate the partial DFT sequence $\{\tilde{x}_k\}_{k=0}^{\frac{N}{2}}$.
2. $\mathbf{Re}[\tilde{x}_k]$ is even since the cosine function is even.
3. $\mathbf{Im}[\tilde{x}_k]$ is odd since the sine function is odd.

3.9.3 Statistical Properties Of The DFT series Of Random Variables. Let $\vec{X} = (X_0, X_1, X_2, \ldots, X_{N-1})$ be a normal vector where N is an even number and each element X_m is an independent normal random variable such that for $m = 0, 1, 2, \ldots, N-1$

$$X_m \sim N(\mu_m, \sigma^2).$$

The mean of the vector $\vec{X}$ is defined as

$$\vec{\mu} = (\mu_0, \mu_1, \mu_2, \ldots, \mu_{N-1}). \tag{3.149}$$

Using equation 3.139, the discrete Fourier transform of $\vec{X}$ is as follows

$$\tilde{X}_k = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} [X_m e^{i(\frac{2\pi}{N})km}]. \tag{3.150}$$

Similarly, the discrete Fourier transform of $\vec{\mu}$ is as follows

$$\tilde{\mu}_k = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} [\mu_m e^{i(\frac{2\pi}{N})km}], \tag{3.151}$$

where $0 \leq k \leq N-1$.

Since equation 3.150 represents a linear combination of independent normal random variables, $\tilde{X}_k$ is also an independent normal complex random variable. Using the results from the complex analysis section, we have the following statistical properties of the DFT complex random variable $\tilde{X}_k$ (12)

a. Mean of the complex variable $\tilde{X}_k$:

$$\begin{aligned} \mathbf{E}[\tilde{X}_k] &= \mathbf{E}\Big[\mathbf{R} \qquad + i\,\mathbf{Im}[\tilde{X}_k]\Big] \\ &= \mathbf{E}\Big[\mathbf{Re}[X_k]\Big] + i\,\mathbf{E}\Big[\mathbf{Im}[\tilde{X}_k]\Big]. \end{aligned} \tag{3.152}$$

1. Using equation 3.143, the expected value of the real part is:

$$\begin{aligned} \mathbf{E}\Big[\mathbf{Re}[\tilde{X}_k]\Big] &= \mathbf{E}\left[\frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} \left[X_m \cos\left[(\frac{2\pi}{N})km\right]\right]\right] \\ &= \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} \left[\mathbf{E}[X_m] \cos\left[(\frac{2\pi}{N})km\right]\right] \\ &= \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} \left[\mu_m \cos\left[(\frac{2\pi}{N})km\right]\right]. \end{aligned} \tag{3.153}$$

2. Using equation 3.144, the expected value of the imaginary part is:

$$\begin{aligned}\mathbf{E}\Big[\mathbf{Im}\big[\tilde{X}_k\big]\Big] &= \mathbf{E}\left[\frac{1}{\sqrt{N}}\sum_{m=0}^{N-1}\left[X_m \sin\left[(\frac{2\pi}{N})km\right]\right]\right] \\ &= \frac{1}{\sqrt{N}}\sum_{m=0}^{N-1}\left[\mathbf{E}[X_m] \sin\left[(\frac{2\pi}{N})km\right]\right] \\ &= \frac{1}{\sqrt{N}}\sum_{m=0}^{N-1}\left[\mu_m \sin\left[(\frac{2\pi}{N})km\right]\right]. \end{aligned} \tag{3.154}$$

We conclude then that

$$\mathbf{E}\big[\tilde{X}_k\big] = \tilde{\mu}_k, \tag{3.155}$$

where $\tilde{\mu}_k$ is k^{th} element of the DFT of $\vec{\mu}$ at the frequency k.

b. Variance of the complex variable $\tilde{X}_k$:

$$\mathbf{Var}\big[\tilde{X}_k\big] = \mathbf{Var}\Big[\mathbf{Re}\big[\tilde{X}_k\big]\Big] + \mathbf{Var}\Big[\mathbf{Im}\big[\tilde{X}_k\big]\Big]. \tag{3.156}$$

1. Variance of the real part:

$$\begin{aligned}\mathbf{Var}\Big[\mathbf{Re}\big[\tilde{X}_k\big]\Big] &= \mathbf{Var}\left[\frac{1}{\sqrt{N}}\sum_{m=0}^{N-1}\left[X_m \cos\left[(\frac{2\pi}{N})km\right]\right]\right] \\ &= \frac{1}{N}\sum_{m=0}^{N-1}\left[\mathbf{Var}[X_m] \cos^2\left[(\frac{2\pi}{N})km\right]\right] \\ &= \frac{1}{N}\sum_{m=0}^{N-1}\left[\sigma^2 \cos^2\left[(\frac{2\pi}{N})km\right]\right] \\ &= \frac{\sigma^2}{N}\sum_{m=0}^{N-1}\left[\cos^2\left[(\frac{2\pi}{N})km\right]\right]. \end{aligned} \tag{3.157}$$

Using the following trigonometric identity:

$$\sum_{r=0}^{N-1}\left[\cos\left[r\alpha\right]\right] = \frac{1}{2} + \frac{\sin\left[(N-\frac{1}{2})\alpha\right]}{2\sin\left[\frac{\alpha}{2}\right]}, \tag{3.158}$$

where $\alpha \neq 2q\pi$ for $q \in \mathbf{Z}$. and the fact that:

$$\cos^2[\beta] = \frac{\cos[2\beta]+1}{2},$$

we can write:

$$\begin{aligned}
\sum_{r=0}^{N-1}\left[\cos^2\left[r\theta\right]\right] &= \sum_{r=0}^{N-1}\left[\frac{\cos[2r\theta]+1}{2}\right] \\
&= \frac{N}{2} + \frac{1}{2}\sum_{r=0}^{N-1}\left[\cos\left[2r\theta\right]\right] \\
&= \frac{N}{2} + \frac{1}{2}\left[\frac{1}{2} + \frac{\sin\left[(N-\frac{1}{2})2\theta\right]}{2\sin\left[\frac{2\theta}{2}\right]}\right] \\
&= \frac{N}{2} + \frac{1}{4}\left[1 + \frac{\sin\left[(N-\frac{1}{2})2\theta\right]}{\sin\left[\theta\right]}\right],
\end{aligned} \tag{3.159}$$

provided that $\theta \neq q\pi$ for $q \in \mathbf{Z}$.

Going back to equation 3.157, we can use equation 3.159 with:

$$\theta = \frac{2\pi k}{N},$$

where $\theta \neq q\pi$ for $q \in \mathbf{Z}$ implies that $k \neq 0$ and $k \neq \frac{N}{2}$. The result is as follows

$$\sum_{r=0}^{N-1}\left[\cos^2\left[r\theta\right]\right] = \frac{N}{2} + \frac{1}{4}\left[1 + \frac{\sin\left[(N-\frac{1}{2})2\theta\right]}{\sin\left[\theta\right]}\right]$$

$$
\begin{aligned}
&= \frac{N}{2} + \frac{1}{4}\left[1 + \frac{\sin\left[(N-\frac{1}{2})2(\frac{2\pi k}{N})\right]}{\sin\left[(\frac{2\pi k}{N})\right]}\right] \\
&= \frac{N}{2} + \frac{1}{4}\left[1 + \frac{\sin\left[(4\pi k - \frac{2\pi k}{N})\right]}{\sin\left[(\frac{2\pi k}{N})\right]}\right] \\
&= \frac{N}{2} + \frac{1}{4}\left[1 - \frac{\sin\left[(\frac{2\pi k}{N})\right]}{\sin\left[(\frac{2\pi k}{N})\right]}\right] \\
&= \frac{N}{2} + \frac{1}{4}\left[1 - 1\right] \\
&= \frac{N}{2}. \qquad (3.160)
\end{aligned}
$$

The variance of the real part of $\tilde{X}_k$ is as follows

1a. When $k \neq 0$ and $k \neq \frac{N}{2}$:

$$
\begin{aligned}
\mathbf{Var}\left[\mathbf{Re}\left[\tilde{X}_k\right]\right] &= \frac{\sigma^2}{N}\sum_{m=0}^{N-1}\left[\cos^2\left[(\frac{2\pi}{N})km\right]\right] \\
&= \frac{\sigma^2}{N}\frac{N}{2} \\
&= \frac{\sigma^2}{2}. \qquad (3.161)
\end{aligned}
$$

1b. When $k = 0$:

$$
\begin{aligned}
\mathbf{Var}\left[\mathbf{Re}\left[\tilde{X}_0\right]\right] &= \frac{\sigma^2}{N}\sum_{m=0}^{N-1}\left[(1)^2\right] \\
&= \frac{\sigma^2}{N}N \\
&= \sigma^2. \qquad (3.162)
\end{aligned}
$$

1c. When $k = \frac{N}{2}$:

$$
\begin{aligned}
\mathbf{Var}\left[\mathbf{Re}\left[\tilde{X}_{\frac{N}{2}}\right]\right] &= \frac{\sigma^2}{N}\sum_{m=0}^{N-1}\left[(-1)^2\right] \\
&= \frac{\sigma^2}{N}N
\end{aligned}
$$

$$= \sigma^2. \tag{3.163}$$

2. Variance of the imaginary part:

$$\begin{aligned}
\mathbf{Var}\Big[\mathbf{Im}[\tilde{X}_k]\Big] &= \mathbf{Var}\left[\frac{1}{\sqrt{N}}\sum_{m=0}^{N-1}\left[X_m \sin\left[(\frac{2\pi}{N})km\right]\right]\right] \\
&= \frac{1}{N}\sum_{m=0}^{N-1}\left[\mathbf{Var}[X_m]\sin^2\left[(\frac{2\pi}{N})km\right]\right] \\
&= \frac{1}{N}\sum_{m=0}^{N-1}\left[\sigma^2 \sin^2\left[(\frac{2\pi}{N})km\right]\right] \\
&= \frac{\sigma^2}{N}\sum_{m=0}^{N-1}\left[\sin^2\left[(\frac{2\pi}{N})km\right]\right] \\
&= \frac{\sigma^2}{N}\sum_{m=0}^{N-1}\left[1-\cos^2\left[(\frac{2\pi}{N})km\right]\right] \\
&= \sigma^2 - \frac{\sigma^2}{N}\sum_{m=0}^{N-1}\left[\cos^2\left[(\frac{2\pi}{N})km\right]\right] \\
&= \sigma^2 - \mathbf{Var}\Big[\mathbf{Re}[\tilde{X}_k]\Big]
\end{aligned} \tag{3.164}$$

3.9.4 Summary Of The Statistics Of The DFT Of Random Variables. Given a normal vector $\vec{X} = (X_0, X_1, X_2, \ldots, X_{N-1})$ and its mean vector $\vec{\mu} = (\mu_0, \mu_1, \mu_2, \ldots, \mu_{N-1})$ where N is an even number and each element X_m is a real independent normal random variable such that for $m = 0, 1, 2, \ldots, N-1$

$$X_m \sim N(\mu_m, \sigma^2),$$

The elements of the DFT of $\vec{X}$ has the following distributions:

Define the mean ,$\tilde{\mu}_k$, of the k^{th} complex coefficient by

$$\tilde{\mu}_k = \frac{1}{\sqrt{N}}\sum_{m=0}^{N-1}\left[\mu_m e^{i(\frac{2\pi}{N})km}\right], \tag{3.165}$$

where μ_m is the mean of the independent normal random variable X_m. we have

a. $k \neq 0$ and $k \neq \frac{N}{2}$

$$\mathbf{Re}[\tilde{X}_k] \sim N(\mathbf{Re}[\tilde{\mu}_k], \frac{\sigma^2}{2}) \tag{3.166}$$

$$\mathbf{Im}[\tilde{X}_k] \sim N(\mathbf{Im}[\tilde{\mu}_k], \frac{\sigma^2}{2}) \tag{3.167}$$

b. $k = 0$ or $k = \frac{N}{2}$

$$\mathbf{Re}[\tilde{X}_k] \sim N(\mathbf{Re}[\tilde{\mu}_k], \sigma^2) \tag{3.168}$$

$$\mathbf{Im}[\tilde{X}_k] = 0. \tag{3.169}$$

IV. Speech De-noising Systems

4.1 Introduction

In this chapter, we present several speech de-noising systems (SDS) using Stein's criteria, wavelets, Fourier, and both the soft thresholding technique (STT) and the hard thresholding technique (HTT). We begin this chapter by an overview of our speech de-noising algorithm, a summary of the main characteristics of speech (voiced, unvoiced, silent, pitch, and formant frequencies), and finally, we present our SDSs.

The speech de-noising systems we developed are applied to noisy voiced portions only. The unvoiced and silent speech portions are processed using a multiplication ratio based on the results of de-noising the voiced portions. Some of our SDSs use the noisy phase in order to eliminate the phase distortions caused by the non-linear processing of the STT and HTT thresholding techniques.

4.2 Speech De-noising Systems Using The SURE Criteria

We present several techniques that are based on using the SURE criteria described in Chapter 3. These techniques assume the following restrictions:

1. A clean speech signal has additive white Gaussian noise which has a normal distribution with zero-mean and variance of σ^2.

2. Only voiced speech is subjected to the de-noising process.

3. Unvoiced speech and the silent portions are not subjected to the de-noising process, instead, they are adjusted by an energy-related ratio to be defined later.

4. The location of the voiced, unvoiced, and silent portions of speech are assumed to be known.

5. The variance required by the SURE function is calculated using an estimate from the silent portions of the speech.

4.2.1 Characteristics Of Speech. In order to understand how the human speech is produced, we are obliged to study and characterize the vocal organs responsible for its production. The vocal organs work by using compressed air which is supplied by the lungs through the trachea (19). The compressed air can then be subjected to periodic pulses (excitations) by the vocal cords (the glottis). The repetition rate of these pulses is termed **pitch** and the resulting periodic speech is termed voiced. When the compressed air passing through the vocal cords is not periodically excited and is forced to passe through a small opening, an air turbulence occurs and a wide-band or broadband noise-like sound is generated. This speech sound is termed unvoiced. After passing through the glottal output, the speech sound, voiced or unvoiced, is subjected to a filtering operation by the shape of the vocal tract. This organ acts as an acoustical tube which strongly passes some natural frequencies which are termed formants.

We conclude then that speech is a signal that is mainly composed of voiced and unvoiced sounds. Voiced speech is characterized by a periodic behavior where the fundamental frequency and the pitch frequency may range from $30Hz$ to about $500Hz$ (19). The pitch varies between males and females. Normally, the pitch frequency is about 125Hz. In our future discussions, we will assume a typical pitch frequency of 125Hz. On the other hand, unvoiced speech has virtually no periodicity and behaves like wide-band noise with less energy than voiced speech. If a speech signal is clean, the energy of the periodic voiced portions is concentrated in bands of frequencies which are harmonics of the fundamental frequency. The pitch frequency, the first, second, and third formant frequencies are normally located below the 3kHz frequency. The energy of the unvoiced portions has a broad-band energy distribution similar to that of noise.

4.2.2 De-noising Algorithm. We developed a speech de-noising algorithm having features described below.

1. The user inputs the following parameters:

   a. The noisy speech file name and the number of samples in this file.

b. The file containing the characteristics of each speech segment: start sample number, end sample number, and status (i.e., voiced, unvoiced. or silent).

c. The number of overlap points between adjacent segments.

d. The percent, p, of the energy of the unvoiced and silent portions to keep.

e. The domain where the de-noising is to take place: time, Fourier (Real and Imaginary), Fourier (Real and Imaginary) to be constructed using noisy phase, wavelets, or any combination of the last four domains.

f. If the wavelets are not involved in the process, the user chooses between using soft or hard thresholding.

g. If the user chooses the wavelet domain, the following parameters are also requested:

i . The wavelet filter and the number of filter points.

ii . The number of decomposition levels.

iii. The thresholding method for the details: soft or hard thresholding.

iv . The de-noising process for the approximation coefficients. The choices include: soft or hard thresholding, no change to the approximations, or energy reduction of the approximations by the same amount as the energy change, R_d , of the processed details.

2. The program searches for the first silent portion and estimates the variances (see equation 4.2).

3. Using the input information from part 1 and the variance from part 2, the program searches for the first voiced portion, multiplies it by a window function using the overlap specified by the user (see equation 4.3), and applies the de-noising process specified by the user.

4. The program calculates the energy ratio R_v between the de-noised voiced portion and the noisy voiced portion.

5. After initializing the variance obtained by step 2 and the energy ratio obtained by step 4, the program steps through the segments file starting from the beginning as follows:

a. read the speech segment and multiply it by a window function using the overlap specified by the user.

b. If the segment is silent:

i . Update the variance.

ii. Multiply this segment by the energy ratio R_v and the percent choice p.

c. If the segment is unvoiced, multiply this segment by the energy ratio R_v and the percent choice p.

d. If the segment is voiced:

i . Apply the de-noising process specified by the user.

ii. Update the energy ratio R_v.

4.2.3 Variance Estimation And The Window Function. The use of the SURE function (see equation 3.28), requires the knowledge of the variance σ^2. Generally, when processing noisy speech data, we don't know in advance the value of this variance. One way of estimating this variance, is to detect the speech silent portions and then use the statistics about white Gaussian noise in order to estimate the variance σ^2.

Given a silent noisy speech portion, $\vec{X} = \{X_i\}_{i=0}^{N-1}$, we estimated the variance using the following consistent estimators as described in (12):

$$\bar{X} = \frac{\sum_{i=0}^{N-1} X_i}{N}, \tag{4.1}$$

the estimate of the variance σ^2 is given by

$$\hat{\sigma}^2 = \frac{\sum_{i=0}^{N-1} (X_i - \bar{X})^2}{N}. \tag{4.2}$$

We mentioned earlier that before processing any speech segment, we multiply it by a window function. In speech processing, it is important to window a speech data before processing it. The reason for using windows is to analyze a finite segment at a time. The length of the window may

vary depending on the desired properties of the signal under analyses (i.e., pitch, time resolution, frequency resolution). However, both the type and the filtering characteristics of the window function play an important role in the results of the analysis. Ideally, we would like a window whose Fourier transform does not have any side-lobe peaks. In practice, we use many different windows, such as the Bartlett window, The Hamming window, and the Hanning window.

Since parts of our algorithm use the discrete wavelet transform (DWT) which is implemented using a periodic extension of the signal under analysis, we chose to implement our window using smooth functions. The trigonometric functions, sines and cosines, are good examples of smooth function. Our window is implemented as follows:

$$win(k) = \begin{cases} \sin^2\left(\frac{2\pi k}{4\delta}\right) & t_b - \frac{\delta}{2} \leq k \leq t_b + \frac{\delta}{2} \\ 1 & t_b + \frac{\delta}{2} < k < t_e - \frac{\delta}{2} \\ 1 - \sin^2\left(\frac{2\pi k}{4\delta}\right) & t_e - \frac{\delta}{2} \leq k \leq t_e + \frac{\delta}{2} \\ 0 & elsewhere, \end{cases} \tag{4.3}$$

where δ is the overlap between adjacent windows (i.e., all our speech experiments have an overlap of $\delta = 16$). The overlap between three adjacent windows are illustrated in figure 4.1. Figure 4.2 illustrates the window described by equation 4.3 and its Fourier transform. Observe that the time domain function has smooth transitions from both ends in order to avoid the introduction of sudden discontinuities caused by a purely rectangular window.

4.2.4 De-noising The Unvoiced And Silent Portions Of Speech. The unvoiced and silent portions of noisy speech have characteristics that are similar to the characteristics of noise. Since the SURE function treats them as white Gaussian noise and tries to eliminate these portions, we decided to de-noise only the voiced portions (see figures D.1 through D.5 for white Gaussian noise and figures D.6 through D.15 for unvoiced speech). The speech without silent and unvoiced portions tends to sound distorted and is hard to understand. For these reasons, we choose not to process the silent and unvoiced portions; instead, we multiply them by the percent ($p = 50\%$) and the energy

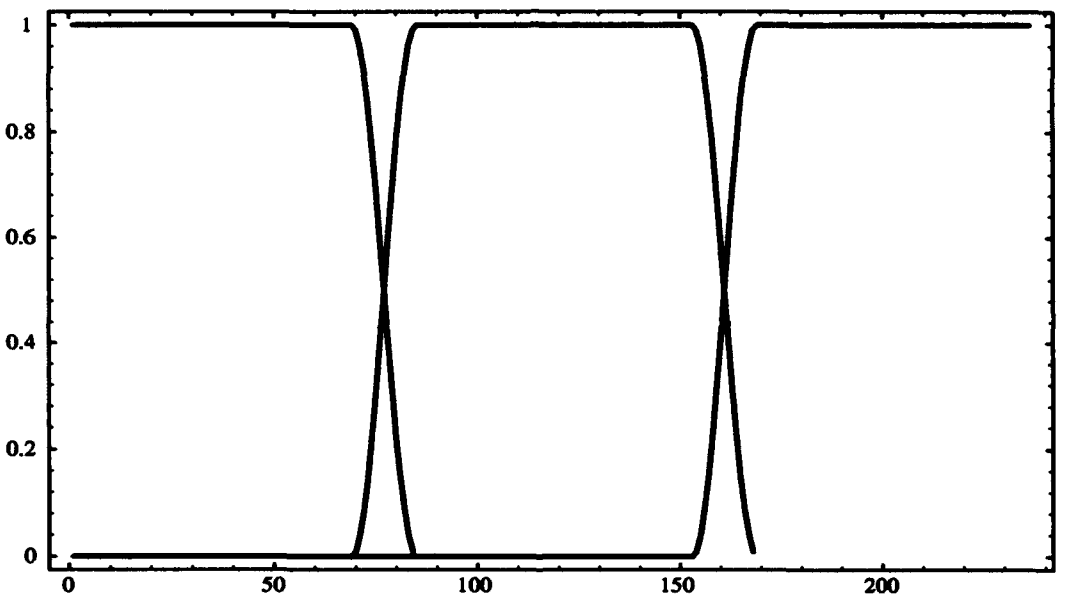

Figure 4.1 Overlap of three window where the overlap $\delta = 16$.

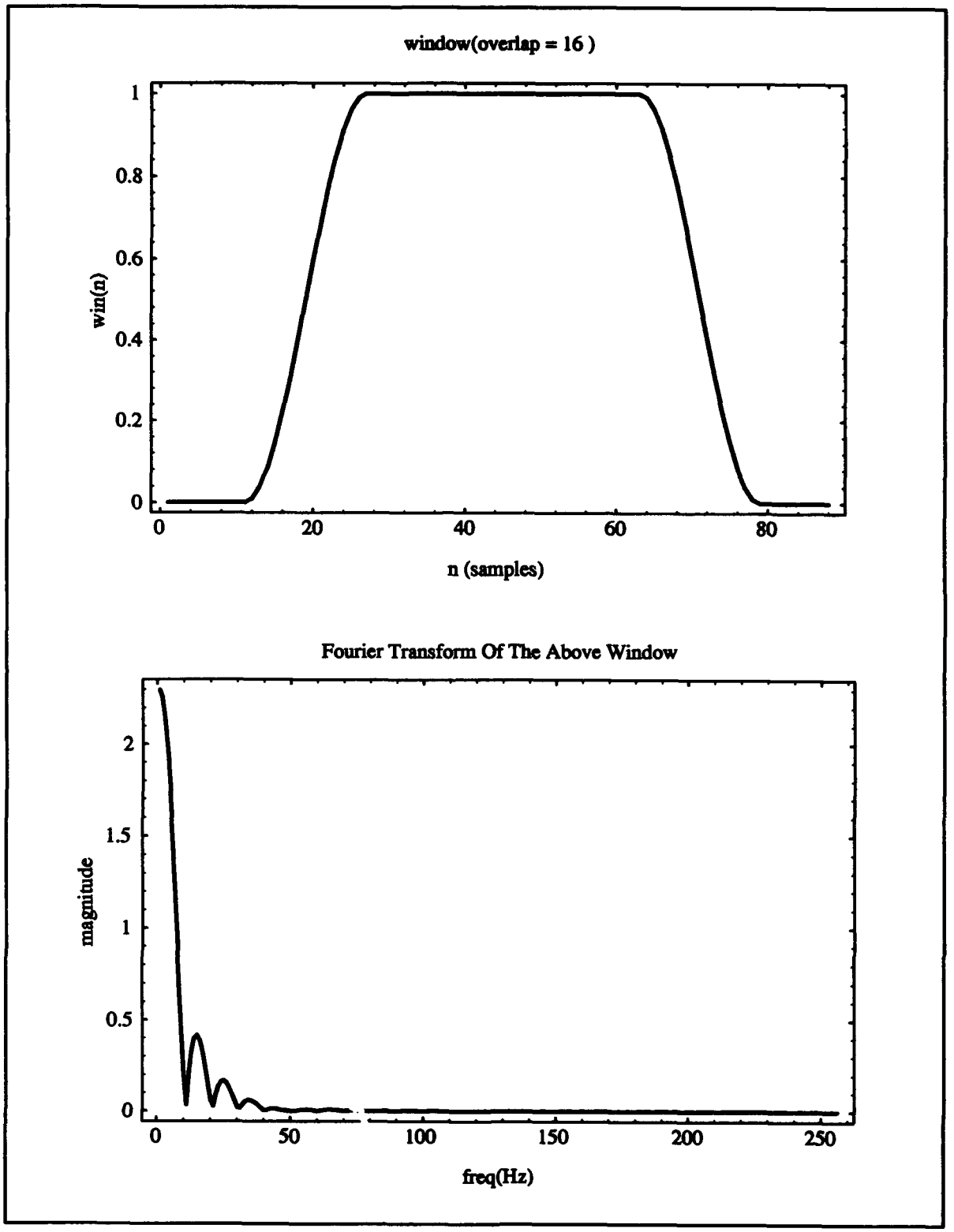

Figure 4.2 Speech window and its Fourier transform.

ratio (R_v) of the change of the energy of the voiced portions as

$$R_v = \sqrt{\frac{E_v^t}{E_v}}, \tag{4.4}$$

where

$$E_v = \sum_{n=0}^{N-1} x_n^2, \tag{4.5}$$

and

$$E_v^t = \sum_{n=0}^{N-1} (x_n^t)^2, \tag{4.6}$$

where x_n and x_n^t are the noisy and thresholded (STT or HTT) voiced speech samples, respectively. Since the noisy samples are thresholded, we have $E_v^t \leq E_v$. The voiced ratio is then constrained as

$$0 \leq R_v \leq 1. \tag{4.7}$$

The new silent and unvoiced samples are then defined as

$$x_n^{(s,new)} = p\, R_v\, x_n^{(s,noisy)} \tag{4.8}$$

$$x_n^{(u,new)} = p\, R_v\, x_n^{(u,noisy)}, \tag{4.9}$$

where $x_n^{(s,noisy)}$, $x_n^{(u,noisy)}$, $x_n^{(s,new)}$, and $x_n^{(u,new)}$ are the silent noisy samples, unvoiced noisy samples, the silent reduced samples, and unvoiced reduced samples, respectively. The ratio R_v helps balance the energy between the voiced, unvoiced, and silent portions, as well as reduce the power of the noise in the silent and unvoiced portions.

4.2.5 De-noising in the time domain. Given a noisy voiced speech signal $\vec{X} = (X_0, X_1, X_2, \ldots, X_{N-1})$ such that

$$\vec{X} = \vec{S} + \vec{Z}, \tag{4.10}$$

where $\vec{S} = (S_0, S_1, S_2, \ldots, S_{N-1})$ is a clean speech vector and $\vec{Z} = (Z_0, Z_1, Z_2, \ldots, Z_{N-1})$ is a white Gaussian noise vector such that for $m = 0, 1, 2, \ldots, N-1$

$$Z_m \sim N(0, \sigma^2),$$

the expected value, $\vec{\mu}$, of the clean speech data $\vec{S}$ is given by

$$\vec{S} = \vec{\mu},$$

where $\vec{\mu} = (\mu_0, \mu_1, \mu_2, \ldots, \mu_{N-1})$.

The noisy vector $\vec{X}$, which is formed by the sum of the constant vector $\vec{S}$ and the normal vector $\vec{Z}$, has a normal distribution with mean $\vec{\mu}$ such that

$$X_m \sim N(\mu_m, \sigma^2), \tag{4.11}$$

where $m = 0, 1, 2, \ldots, N-1$.

Since $\vec{X}$ has a normal distribution, we can directly use the time domain speech data degraded by white Gaussian noise with the SURE function (see figure 4.3). The time domain speech de-noising system (SDS) has the advantage of not requiring further transformations which are time consuming. However, the application of either the soft or the hard thresholding techniques to a segment of speech in the time domain, uses a single threshold to adjust a whole window of speech.

This threshold may not be sufficient to eliminate most of the noise and hence we may expect that the output of the time SDS to be only slightly cleaner than the input speech.

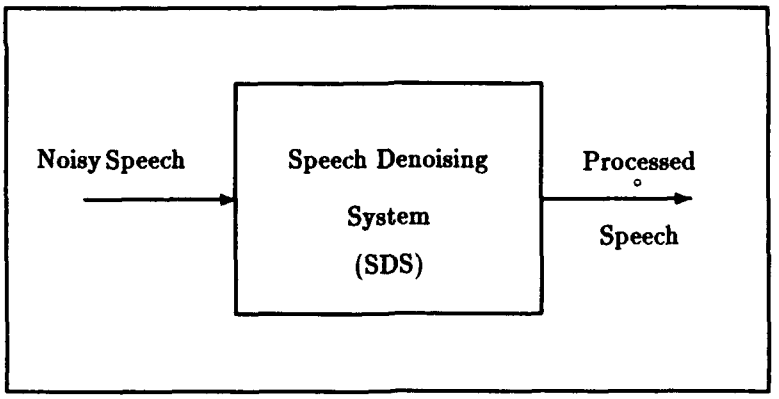

Figure 4.3 Speech de-noising in the time domain

4.2.6 De-noising in the time domain using the noisy phase. In the speech processing field, it is believed that some of the distortion caused by de-noising speech data is mainly due to the change of the phase in the Fourier representation of speech. These distortions may diminish the intelligibility of the de-noised speech. In order to study the effect of the phase, θ, on our SDS and on the intelligibility of the de-noised speech, we save the noisy phase for reconstruction and apply the de-noising techniques described in the previous section (see figure 4.4).

Although this technique improves intelligibility, it requires more processing due to the Fourier transform and more storage due to phase saving.

4.2.7 De-noising in the frequency domain. We have seen that if the real and imaginary parts of a complex random variable are normal, the amplitude and phase can't be normal. Since the Fourier transform is a linear operation, the Fourier transform of a normal multivariate vector is also normal. However, the variance of the Fourier transform coefficients were shown to be not identical (e.g., dc component). Recall the discrete Fourier transform (DFT) of a periodic finite-

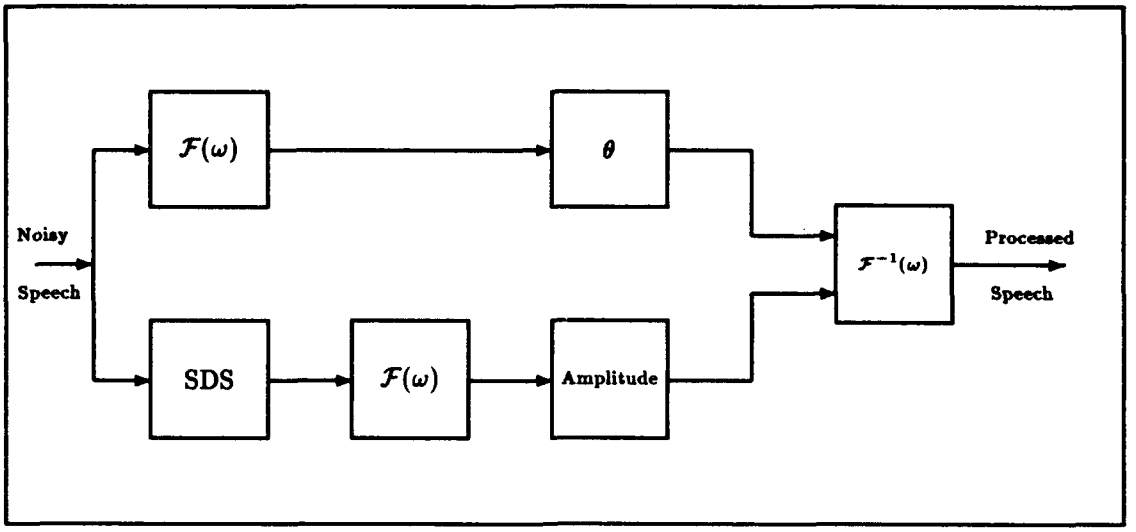

Figure 4.4 Speech de-noising in the time domain using noisy phase

length sequence of N points, $\{X_m\}_{m=0}^{N-1}$ is

$$\tilde{X}_k = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} \left[X_m e^{i(\frac{2\pi}{N})km} \right], \tag{4.12}$$

where $0 \leq k \leq N-1$. We have shown that if the time sequence $\{X_m\}_{m=0}^{N-1}$ has a normal distribution such that each random variable $X_m \sim N(\mu_m, \sigma^2)$, the Fourier sequence $\{\tilde{X}_k\}_{k=0}^{N-1}$ is conjugate symmetric such that the real and imaginary parts of $\tilde{X}_k$ have the normal distributions $N(\mathbf{Re}[\tilde{\mu}_k], \frac{\sigma^2}{2})$ and $N(\mathbf{Im}[\tilde{\mu}_k], \frac{\sigma^2}{2})$ for $1 \leq k < \frac{N}{2}$, respectively. However, the 0^{th} and the $\frac{N}{2}$ real and imaginary elements are distributed according to $N(\mathbf{Re}[\tilde{\mu}_k], \sigma^2)$ and $N(\mathbf{Im}[\tilde{\mu}_k], 0)$, respectively. This property of the Fourier coefficients allows us to use the sequence of real and imaginary elements 1 through $(\frac{N}{2} - 1)$, inclusive, with the SURE function which requires the input random variables to be normal, independent, and to have the same variance.

The method calls for processing separately, the two time sequences $\left\{\mathbf{Re}[\tilde{X}_k]\right\}_{k=1}^{(\frac{N}{2}-1)}$ and $\left\{\mathbf{Im}[\tilde{X}_k]\right\}_{k=1}^{(\frac{N}{2}-1)}$, where each element has a normal distribution with variance $\frac{\sigma^2}{2}$ (see figure 4.5). After the application of the SURE threshold, the real and imaginary outputs are combined with the original dc component and the $\frac{N}{2}$ component and then inverse

Fourier transformed to produce back the time domain de-noised signal. The elements $\{\tilde{X}_0, \tilde{X}_{\frac{N}{2}}\}$ are left untouched because of their unique distributions and characteristics (see equations 3.168 and 3.169). Depending on how the DFT is defined, the dc component, $\tilde{X}_0$, is a measure of the mean of the time sequence $\{X_m\}_{m=0}^{N-1}$, while $\tilde{X}_{\frac{N}{2}}$ is the high frequency component. Since noise is generally composed of high frequencies, little or no modification to the dc component may occur.

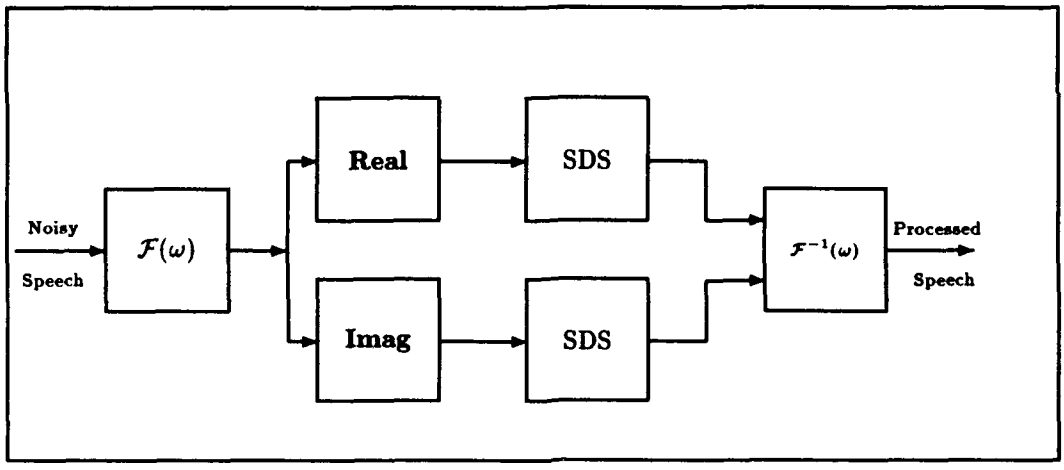

Figure 4.5 Speech de-noising in the frequency domain

4.2.7.1 Soft Thresholding Of Complex Data. When using the shrinkage or the soft thresholding technique (STT) in the Fourier domain, the real and imaginary parts are affected in a way that affects the phase of the complex Fourier coefficients being de-noised. Consider the k^{th} Fourier coefficient where $k = 1, 2, \ldots, (\frac{N}{2} - 1)$, and denote the real and imaginary soft thresholds by t_s^{Re} and t_s^{Im}, respectively. Because of the definition of the STT, which pulls a noisy data sample towards zero if its magnitude is greater than the threshold or sets it to zero otherwise, we have four different cases (see figure 4.6).

Define the new modified complex number, $\tilde{X}_k^{t,soft}$, by

$$\tilde{X}_k^{t,soft} = \mathbf{Re}[\tilde{X}_k^{t,soft}] + i\,\mathbf{Im}[\tilde{X}_k^{t,soft}], \tag{4.13}$$

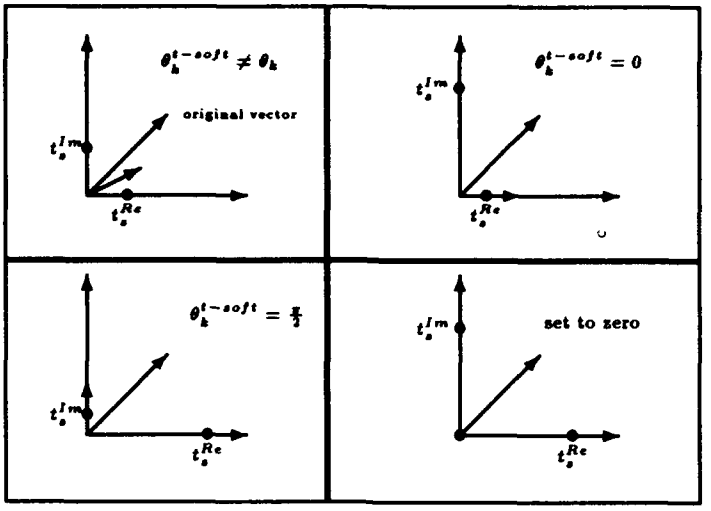

Figure 4.6 Four possible changes and orientations of a de-noised complex number using the STT.

where the de-noised real part is defined as (see equation 3.32)

$$\mathbf{Re}[\tilde{X}_k^{t,soft}] = \mathbf{Re}[\tilde{X}_k] - \min\left(\left|\mathbf{Re}[\tilde{X}_k]\right|, t_s^{Re}\right) \operatorname{sgn}\left(\mathbf{Re}[\tilde{X}_k]\right), \tag{4.14}$$

and the imaginary part is defined as

$$\mathbf{Im}[\tilde{X}_k^{t,soft}] = \mathbf{Im}[\tilde{X}_k] - \min\left(\left|\mathbf{Im}[\tilde{X}_k]\right|, t_s^{Im}\right) \operatorname{sgn}\left(\mathbf{Im}[\tilde{X}_k]\right). \tag{4.15}$$

Combining equation 4.14 and 4.15 we get

$$\begin{aligned}
\tilde{X}_k^{t,soft} &= \mathbf{Re}[\tilde{X}_k^{t,soft}] + i\,\mathbf{Im}[\tilde{X}_k^{t,soft}] \\
&= \left[\mathbf{Re}[\tilde{X}_k] - \min\left(\left|\mathbf{Re}[\tilde{X}_k]\right|, t_s^{Re}\right) \operatorname{sgn}\left(\mathbf{Re}[\tilde{X}_k]\right)\right] + \\
&\quad i\left[\mathbf{Im}[\tilde{X}_k] - \min\left(\left|\mathbf{Im}[\tilde{X}_k]\right|, t_s^{Im}\right) \operatorname{sgn}\left(\mathbf{Im}[\tilde{X}_k]\right)\right] \\
&= \left[\mathbf{Re}[\tilde{X}_k] + i\,\mathbf{Im}[\tilde{X}_k]\right] - \\
&\quad \left[\min\left(\left|\mathbf{Re}[\tilde{X}_k]\right|, t_s^{Re}\right) \operatorname{sgn}\left(\mathbf{Re}[\tilde{X}_k]\right) + i \min\left(\left|\mathbf{Im}[\tilde{X}_k]\right|, t_s^{Im}\right) \operatorname{sgn}\left(\mathbf{Im}[\tilde{X}_k]\right)\right] \\
&= \tilde{X}_k + g^{t,soft}[\tilde{X}_k],
\end{aligned} \tag{4.16}$$

where

$$g^{t,soft}[\tilde{X}_k] = -\min\left(\left|\mathbf{Re}[\tilde{X}_k]\right|, t_s^{Re}\right)\operatorname{sgn}\left(\mathbf{Re}[\tilde{X}_k]\right) - i\,\min\left(\left|\mathbf{Im}[\tilde{X}_k]\right|, t_s^{Im}\right)\operatorname{sgn}\left(\mathbf{Im}[\tilde{X}_k]\right). \tag{4.17}$$

The above $g^{t,soft}$ function is the complex equivalent of the real $g_i^{t,soft}(\vec{X})$ function defined by equation 3.31. The phase (provided it exists) of the de-noised complex coefficient, $\tilde{X}_k^{t,soft}$, is defined as

$$\begin{aligned}\theta_k^{t,soft} &= \arctan\left[\frac{\mathbf{Im}[\tilde{X}_k^{t,soft}]}{\mathbf{Re}[\tilde{X}_k^{t,soft}]}\right] \\ &= \arctan\left[\frac{\mathbf{Im}[\tilde{X}_k] - \min\left(\left|\mathbf{Im}[\tilde{X}_k]\right|, t_s^{Im}\right)\operatorname{sgn}\left(\mathbf{Im}[\tilde{X}_k]\right)}{\mathbf{Re}[\tilde{X}_k] - \min\left(\left|\mathbf{Re}[\tilde{X}_k]\right|, t_s^{Re}\right)\operatorname{sgn}\left(\mathbf{Re}[\tilde{X}_k]\right)}\right].\end{aligned} \tag{4.18}$$

On the other hand, the phase θ_k of the noisy coefficient $\tilde{X}_k$ is defined as

$$\theta_k = \arctan\left[\frac{\mathbf{Im}[\tilde{X}_k]}{\mathbf{Re}[\tilde{X}_k]}\right]. \tag{4.19}$$

We see from equations 4.18 and 4.19 that this new shrinkage technique applied to the real and imaginary parts separately, has the potential to introduce a lot of distortion due to the phase changes of the entire frequency spectrum. In fact when the thresholds act on the real and imaginary parts, the phase can take any value within its domain (see case $\theta_k^{t,soft} \neq \theta_k$ in figure 4.6). One way of avoiding more phase distortion than present in the noisy signal is to keep the original noisy phase and use it in the inverse Fourier transform back to the time domain.

4.2.7.2 Hard Thresholding Of Complex Data. When using the Hard Thresholding Technique (HTT) in the Fourier domain, the real and imaginary parts are also affected in a way that affects the phase of the complex Fourier coefficients being de-noised. Consider the k^{th} Fourier coefficient where $k = 1, 2, \ldots, (\frac{N}{2} - 1)$, and denote the real and imaginary hard thresholds by t_h^{Re} and t_h^{Im}, respectively. Because of the definition of the HTT, which sets a noisy data sample to zero if its magnitude is less than the threshold , we have four different cases (see figure 4.7).

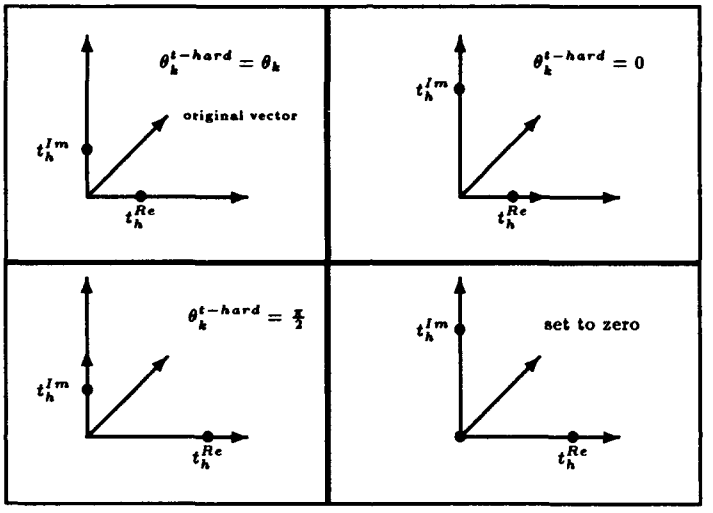

Figure 4.7 Four possible changes and orientations of a de-noised complex number using the HTT.

Define the new modified complex number, $\tilde{X}_k^{t,hard}$, by

$$\tilde{X}_k^{t,hard} = \mathbf{Re}[\tilde{X}_k^{t,hard}] + i\,\mathbf{Im}[\tilde{X}_k^{t,hard}], \tag{4.20}$$

where the de-noised real part is defined as (see equation 3.45)

$$\mathbf{Re}[\tilde{X}_k^{t,hard}] = \mathbf{Re}[\tilde{X}_k]\,\chi_{[-t_h^{Re}, t_h^{Re}]}\Big(\mathbf{Re}[\tilde{X}_k]\Big), \tag{4.21}$$

and the imaginary part is defined as

$$\mathbf{Im}[\tilde{X}_k^{t,hard}] = \mathbf{Im}[\tilde{X}_k]\,\chi_{[-t_h^{Im}, t_h^{Im}]}\Big(\mathbf{Im}[\tilde{X}_k]\Big). \tag{4.22}$$

Combining equation 4.21 and 4.22 we get

$$\begin{aligned}
\tilde{X}_k^{t,hard} &= \mathbf{Re}\left[\tilde{X}_k^{t,hard}\right] + i\,\mathbf{Im}\left[\tilde{X}_k^{t,hard}\right] \\
&= \mathbf{Re}\left[\tilde{X}_k\right] \chi_{[-t_h^{Re}, t_h^{Re}]}\left(\mathbf{Re}\left[\tilde{X}_k\right]\right) + i\,\mathbf{Im}\left[\tilde{X}_k\right] \chi_{[-t_h^{Im}, t_h^{Im}]}\left(\mathbf{Im}\left[\tilde{X}_k\right]\right) \\
&= \tilde{X}_k + g^{t,hard}\left[\tilde{X}_k\right], \qquad (4.23)
\end{aligned}$$

where

$$g^{t,hard}\left[\tilde{X}_k\right] = -\mathbf{Re}\left[\tilde{X}_k\right]\left(1 - \chi_{[-t_h^{Re}, t_h^{Re}]}\left(\mathbf{Re}\left[\tilde{X}_k\right]\right)\right) - i\,\mathbf{Im}\left[\tilde{X}_k\right]\left(1 - \chi_{[-t_h^{Im}, t_h^{Im}]}\left(\mathbf{Im}\left[\tilde{X}_k\right]\right)\right). \qquad (4.24)$$

The above $g^{t,hard}$ function is the complex equivalent to the real $g_i^{t,hard}(\vec{X})$ function defined by equation 3.46. The phase (provided it exists) of the de-noised complex coefficient, $\tilde{X}_k^{t,hard}$, is defined as

$$\begin{aligned}
\theta_k^{t,hard} &= \arctan\left[\frac{\mathbf{Im}\left[\tilde{X}_k^{t,hard}\right]}{\mathbf{Re}\left[\tilde{X}_k^{t,hard}\right]}\right] \\
&= \arctan\left[\frac{\mathbf{Im}\left[\tilde{X}_k\right] \chi_{[-t_h^{Im}, t_h^{Im}]}\left(\mathbf{Im}\left[\tilde{X}_k\right]\right)}{\mathbf{Re}\left[\tilde{X}_k\right] \chi_{[-t_h^{Re}, t_h^{Re}]}\left(\mathbf{Re}\left[\tilde{X}_k\right]\right)}\right] \\
&= \arctan\left[\tan[\theta_k] \frac{\chi_{[-t_h^{Im}, t_h^{Im}]}\left(\mathbf{Im}\left[\tilde{X}_k\right]\right)}{\chi_{[-t_h^{Re}, t_h^{Re}]}\left(\mathbf{Re}\left[\tilde{X}_k\right]\right)}\right], \qquad (4.25)
\end{aligned}$$

where the phase θ_k of the noisy coefficient $\tilde{X}_k$ is defined as

$$\theta_k = \arctan\left[\frac{\mathbf{Im}\left[\tilde{X}_k\right]}{\mathbf{Re}\left[\tilde{X}_k\right]}\right]. \qquad (4.26)$$

We see from equations 4.25 and 4.26 that when the HTT is applied to the real and imaginary parts separately, it has the potential to introduce a lot of distortion due to the phase changes of the entire frequency spectrum. However, these phase distortions can take only four different forms:

1. Don't change the phase.
2. set the phase to zero.
3. set the phase to $\frac{\pi}{2}$.
4. set the noisy data to zero, changing the phase from defined to undefined.

4.2.8 Speech de-noising in the frequency domain using noisy phase. It was noted in the previous section that without saving the noisy phase, we might introduce many phase distortions to the speech signal. In order to improve intelligibility, we save the noisy phase and use the same thresholding process as before (see figure 4.8). In order to restore the noisy phase θ, we need to first apply the thresholding technique as in the previous section, calculate the amplitude of the modified Fourier coefficients, and then combine the amplitude with the noisy phase.

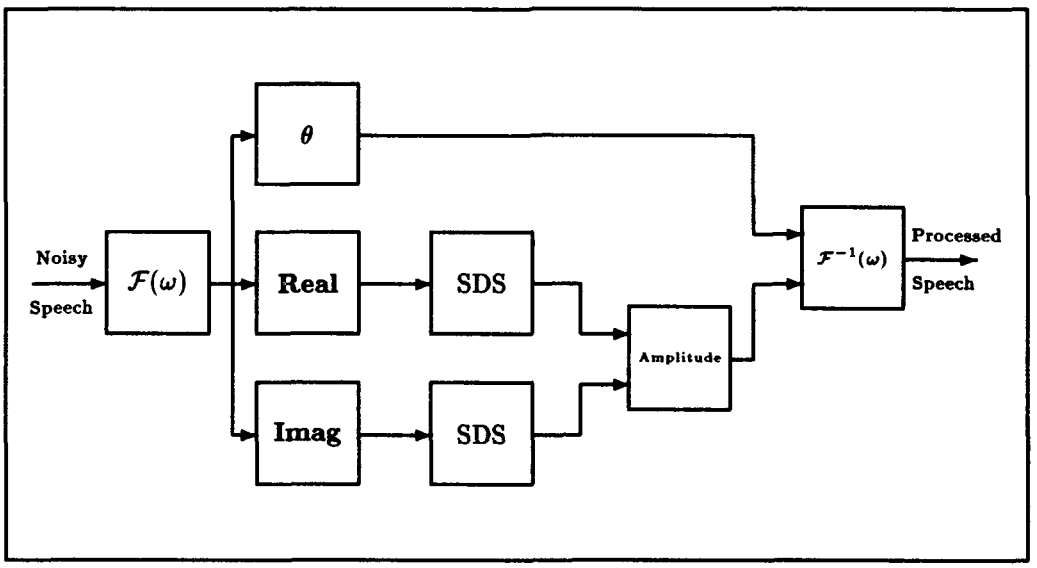

Figure 4.8 Speech de-noising in the frequency domain using noisy phase

4.2.8.1 Soft Thresholding Of Complex Data With Noisy Phase Restoration. Consider the k^{th} Fourier coefficient where $k = 1, 2, \ldots, (\frac{N}{2} - 1)$, and denote the real and imaginary soft thresholds by t_s^{Re} and t_s^{Im}, respectively. Define the modified complex coefficient by the shrinkage technique (see equation 4.13) as

$$\tilde{X}_k^{t,soft} = \mathbf{Re}[\tilde{X}_k^{t,soft}] + i\,\mathbf{Im}[\tilde{X}_k^{t,soft}], \tag{4.27}$$

and denote the new modified complex coefficient with noisy phase restoration by

$$\tilde{X}_k^{t,soft-\theta} = |\tilde{X}_k^{t,soft}|\, e^{\theta_k}, \tag{4.28}$$

where phase θ_k is defined by equation 4.19 and the real and imaginary components of $\tilde{X}_k^{t,soft}$, are as defined by equations 4.14, 4.15 , respectively. In rectangular form, we have

$$\tilde{X}_k^{t,soft-\theta} = \mathbf{Re}[\tilde{X}_k^{t,soft-\theta}] + i\,\mathbf{Im}[\tilde{X}_k^{t,soft-\theta}]. \tag{4.29}$$

Expanding equation 4.28, the new de-noised real part is defined as

$$\mathbf{Re}[\tilde{X}_k^{t,soft-\theta}] = |\tilde{X}_k^{t,soft}|\, \cos(\theta_k), \tag{4.30}$$

and the imaginary part is defined as

$$\mathbf{Im}[\tilde{X}_k^{t,soft-\theta}] = |\tilde{X}_k^{t,soft}|\, sin(\theta_k). \tag{4.31}$$

This new shrinkage technique takes advantage of the normal distribution properties of the real and imaginary parts in order to shrink the amplitude. Pictorially, there are four different cases that we need to consider (see figure 4.9). When applied to the complex number $\tilde{X}_k$, this new

shrinkage technique has no effect on the phase (since the noisy phase is restored), however, the amplitude is affected in one of two different manners:

1. The amplitude is shrunken toward zero by a nonzero amount.
2. The amplitude is set to zero.

We see then that there are a lot of advantages to keeping the noisy phase so that when we inverse Fourier transform, many of the potential phase distortions due to the thresholding techniques are eliminated.

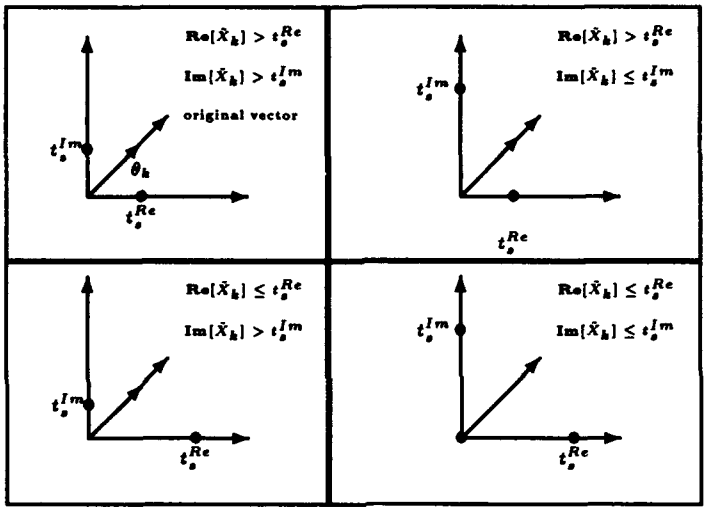

Figure 4.9 Four possible changes and orientations of a de-noised complex number with noisy phase restoration.

4.2.8.2 Hard Thresholding Of Complex Data With Noisy Phase Restoration. Following the same procedure as before and denoting the modified complex coefficient by the hard thresholding technique (see equation 4.20) as

$$\tilde{X}_k^{t,hard} = \mathbf{Re}\left[\tilde{X}_k^{t,hard}\right] + i\,\mathbf{Im}\left[\tilde{X}_k^{t,hard}\right], \tag{4.32}$$

and the new modified complex coefficient with noisy phase restoration by

$$\tilde{X}_k^{t,hard-\theta} = \left|\tilde{X}_k^{t,hard}\right| e^{\theta_k}, \tag{4.33}$$

we obtain the same results as the shrinkage technique. In fact all the equations of the HTT are the same as the STT except for the naming designators (soft and hard). Again four cases can be considered (see figure 4.9).

4.3 Application Of SURE To DWT

We have seen that the wavelet transform is a linear operator and that the detail coefficients, at a decomposition level m, measure the degree of similarity between the signal $f(t)$ and the the analyzing wavelet $\psi_{m,n}(t)$; furthermore, the details give us some degree of information concerning the frequency content of the signal $f(t)$. Recall, due to the down-sampling and filtering operations performed during the decomposition process, the lower levels (i.e., $m = 1, 2, ..$) represent high frequency information, and the higher levels (i.e., $m = M, M-1, M-2, ..$) represent low frequency information.

Define a noisy vector $\vec{X} = (X_0, X_1, X_2, \ldots, X_{N-1})$ such that

$$\vec{X} = \vec{S} + \vec{Z}, \tag{4.34}$$

where $\vec{S} = (S_0, S_1, S_2, \ldots, S_{N-1})$ is a clean data vector and $\vec{Z} = (Z_0, Z_1, Z_2, \ldots, Z_{N-1})$ is a white Gaussian noise vector such that for $m = 0, 1, 2, \ldots, N-1$

$$Z_m \sim N(0, \sigma^2).$$

Since $\vec{S}$ is a constant clean data vector, the expected value of this vector is the vector $\vec{\mu}$ such that

$$\vec{S} = \vec{\mu},$$

where $\vec{\mu} = (\mu_0, \mu_1, \mu_2, \ldots, \mu_{N-1})$.

The noisy vector $\vec{X}$, which is formed by the sum of a constant vector $\vec{S}$ and a normal vector $\vec{Z}$, has

a normal distribution with mean $\vec{\mu}$ such that

$$X_m \sim N(\mu_m, \sigma^2), \tag{4.35}$$

where $m = 0, 1, 2, \ldots, N-1$.

4.3.1 Voiced speech vs. White Gaussian Noise. The wavelet decomposition of the normal random vector $\vec{X}$ at the m^{th}-level ($1 \leq m \leq M$) is given by equations 3.106 and 3.107 as

$$C^X_{m,n} = \sum_{k \in \mathbf{Z}} C^X_{m-1,k} h_{k-2n} \tag{4.36}$$

$$D^X_{m,n} = \sum_{k \in \mathbf{Z}} C^X_{m-1,k} g_{k-2n}, \tag{4.37}$$

where C^X and D^X denote the approximation and detail random variables with respect to $\vec{X}$, respectively. Since the DWT is linear and orthogonal, we have

$$C^X_{m,n} = C^S_{m,n} + C^Z_{m,n} \tag{4.38}$$

$$D^X_{m,n} = D^S_{m,n} + D^Z_{m,n}. \tag{4.39}$$

Using the above results and the fact that the DWT coefficients are also independent and normally distributed, it can easily be shown that the wavelet coefficients (details and approximations) of the white Gaussian noise, $\vec{Z}$, at the m^{th} decomposition level, are themselves white Gaussian noise with zero-mean and the same variance, σ^2. This normal distribution property of the wavelet coefficients makes them candidates for use with the SURE function developed earlier. Since the detail coefficients measure the amount of some frequencies in a well defined band of frequencies (depending on the decomposition level m and the analyzing wavelet $\psi_{m,n}$), we can directly apply the de-noising process to certain bands of frequencies where the white Gaussian noise has a high probability of residing. Since the formant frequencies of voiced speech are relatively low-frequencies

(below 3kHz), and white Gaussian noise uniformly contains all frequencies , the early stages of decomposition have a high probability of filtering most of the high frequencies that are due to noise, while the later stages of decomposition filter the voiced speech signal (see figures E.1 through E.10 for voiced speech de-noising using shrinkage).

Since both the STT and the HTT techniques are non-linear thresholding techniques, we decided to DWT (discrete wavelet transform) our signal up to a decomposition level where the pitch frequency is not affected by the non-linear thresholding (Note: our algorithm gives you an option to process both the approximations and the details). Recall that the DWT is a filtering operation that uses a low-pass filter (h) and a high-pass filter (g). At each level of decomposition, the high-pass g filter divides the frequency spectrum by half. Given a noisy voiced speech signal where the pitch frequency f_p is known and a sampling frequency $f_s = 16kHz$, the maximum resolvable frequency is $\frac{f_s}{2} = 8kHz$ (14) (17). In order not to affect the pitch frequency, we need to decompose up to a level $m \leq m_v$ where

$$m_v = \left\lfloor \log_2 \left[\frac{f_s}{2f_p} \right] \right\rfloor, \tag{4.40}$$

where $\lfloor . \rfloor$ is the floor function. Since our speech data is sampled at 16kHz and we are assuming a typical pitch frequency of 125Hz, the m_v value is 6. By decomposing the signal up to the m_v^{th}-level and applying our thresholding techniques, we have a high chance of eliminating most of the noise in the first m_v levels without affecting the pitch of the voiced speech which resides in the remaining coarser levels. This partial wavelet decomposition of the voiced speech signal yields voiced speech where the structure of the pitch is not subjected to the thresholding techniques, i.e., the pitch is contained mostly at the approximation levels (see figure 4.10).

4.3.2 Wavelet Coefficients Thresholding. Having determined the maximum level of decomposition, m_v, we can apply either the soft thresholding technique or the hard thresholding

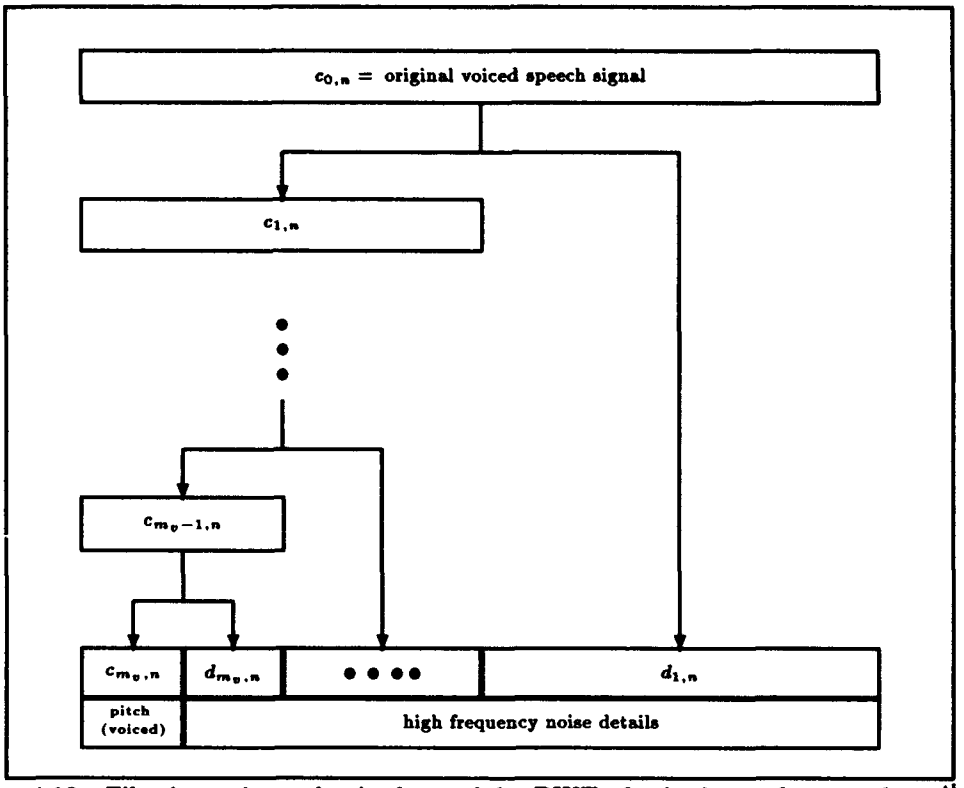

Figure 4.10 Filtering noise and voiced speech by DWT of voiced speech up to the m_v^{th}-level.

technique to each of the m_v levels of details. Consider the noisy signal $\vec{X}$ with $N = 2^M$ points. We know that by applying the DWT discussed earlier, the total number of decomposition levels is M. Define the detail coefficients of the m^{th} decomposition level by $D_{m,n}$ where $1 \leq m \leq m_v$ and $0 \leq n \leq 2^{M-m} - 1$.

4.3.2.1 Wavelet Shrinkage Of The Detail Coefficients. Consider the m^{th} decomposition level and denote the soft threshold at this level by, t_s^m. Define the shrunken version of the detail coefficient $D_{m,n}$ by

$$D_{m,n}^{t_s^m} = D_{m,n} - \min\left(|D_{m,n}|, t_s^m\right) \operatorname{sgn}(D_{m,n}), \tag{4.41}$$

where $0 \leq n < 2^{M-m}$. We can then define a function $g_n^{t_s^m}$ which is level dependent, such that

$$g_n^{t_s^m}(\vec{D}_m) = -\min\left(|D_{m,n}|, t_s^m\right) \operatorname{sgn}(D_{m,n}), \tag{4.42}$$

where $\vec{D}_m$ is the vector whose elements are the 2^{M-m} detail coefficients at the m^{th} decomposition level. Since we have m_v levels, we have m_v thresholds t_s^m, and m_v functions $g_n^{t_s^m}(\vec{D}_m)$.

4.3.2.2 Hard Thresholding Of The Wavelet Detail Coefficients. In a similar fashion, we can apply the hard thresholding technique to the wavelet details and the results are similar to the shrinkage case. Consider the m^{th} decomposition level and denote the hard threshold at this level by, t_h^m. Define the hard thresholded version of the detail coefficient $D_{m,n}$ by

$$D_{m,n}^{t_h^m} = D_{m,n}\, \chi_{[-t_h^m, t_h^m]}(D_{m,n}), \tag{4.43}$$

where $0 \leq n < 2^{M-m}$. We can also define a function $g_d^{t_h^m}$ that is level dependent, such that

$$g_n^{t_h^m}(\vec{D}_m) = -D_{m,n}\left(1 - \chi_{[-t_h^m, t_h^m]}(D_{m,n})\right), \tag{4.44}$$

where $\vec{D}_m$ is the vector whose elements are the 2^{M-m} detail coefficients at the m^{th} decomposition level. Since we have m_v levels, we have m_v thresholds t_h^m, and m_v functions $g_n^{t_h^m}(\vec{D}_m)$.

4.3.2.3 De-noising The Approximations. Since the pitch of the voiced speech is represented by the approximation coefficients at the m_v^{th} decomposition level. The total number of these coefficients is 2^{M-m_v}. In order to prevent this voiced signal from being distorted, we choose to either leave the approximation coefficients $\left\{C_{m_v,n}\right\}_{n=0}^{2^{M-m_v}-1}$ untouched or adjust their energy by the same amount as the energy change of all the thresholded details (STT or HTT). In other

words, the ratio between the energies of the noisy details and the de-noised details is defined by

$$R_d = \sqrt{\frac{E_D^t}{E_D}}, \tag{4.45}$$

where

$$E_D = \sum_{m=1}^{m_v} \left[\sum_{n=0}^{2^{M-m}-1} [D_{m,n}]^2 \right], \tag{4.46}$$

and

$$E_D^t = \sum_{m=1}^{m_v} \left[\sum_{n=0}^{2^{M-m}-1} [D_{m,n}^t]^2 \right]. \tag{4.47}$$

Since the noisy details are thresholded, we have $E_D^t \leq E_D$. The detail ratio is then constrained as

$$0 \leq R_d \leq 1. \tag{4.48}$$

The new approximation coefficients at the m_v^{th}-level are then defined as

$$C_{m_v,n}^t = R_d \, C_{m_v,n}, \tag{4.49}$$

where $0 \leq n < 2^{M-m_v}$. The ratio R_p helps balance the energy between the approximations and the details as well as reduce the power of any noise that passed through the m_v decomposition level (see figure 4.11).

4.3.3 De-noising The DWT of The Time Domain. We have seen that the the wavelet transform of a normal multidimensional random vector, produces a set of detail coefficients vectors that are also normal. By applying the SURE thresholding techniques to these details (see figure 4.12), we can eliminate most of the noise at the first m_v levels. Since the wavelets are band-pass

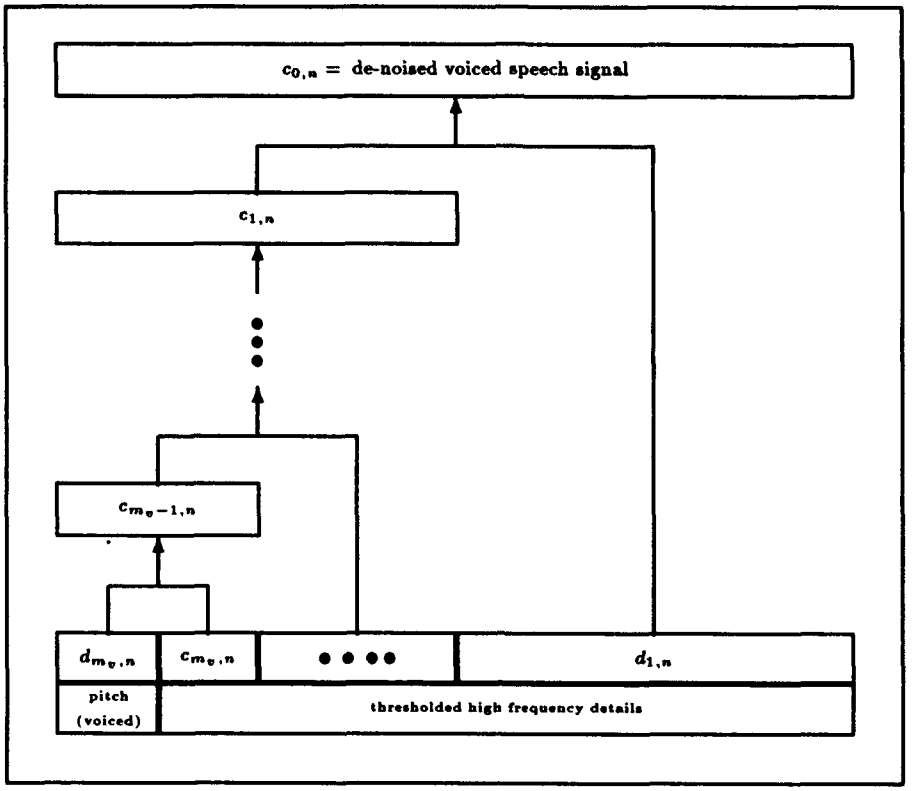

Figure 4.11 Wavelet reconstruction of the thresholded (STT or HTT) voiced speech starting from the m_v^{th}-level where $1 \leq m_v \leq M$ to the *zeroth* level where the number of samples is $N = 2^M$

filters, at each level of decomposition, an entire band of high frequencies is being de-noised. We expect then that the output of this method to eliminate most of the high frequencies that are mainly due to noise.

A variation of this purely time-wavelet domain scheme may be employed to minimize the phase distortion introduced by the nonlinear effect of the thresholding techniques (STT and HTT). In order to reduce the effect of phase distortions, we may save the noisy phase from the Fourier transform of the noisy voiced speech and restore it after the de-noising procedures. Figure 4.13 illustrates the method; the time domain voiced speech waveform is first Fourier transformed to extract the phase and then wavelet transformed before the de-noising process is applied. The thresholded details, are then inverse wavelet transformed, Fourier transformed in order to extract the de-noised amplitude. Finally, the old phase is combined with this newly calculated amplitude

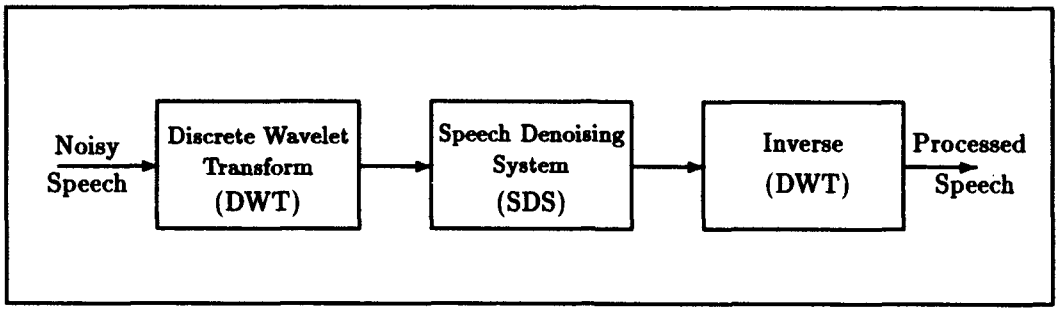

Figure 4.12 Speech de-noising in the time domain using wavelets

and inverse Fourier transformed back to the time domain. Observe, this method requires three Fourier transforms and two wavelet transforms.

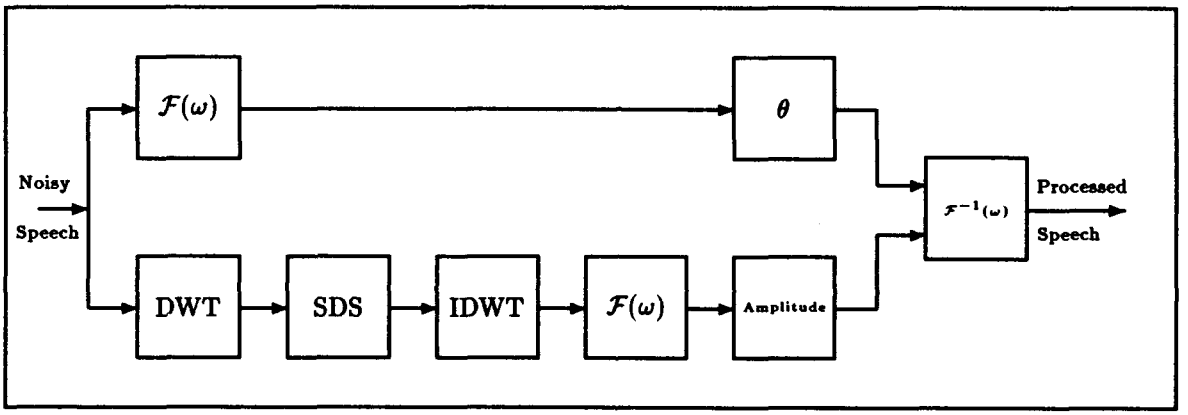

Figure 4.13 Speech de-noising in the time domain using noisy phase and wavelets

4.3.4 De-noising The DWT of The Fourier Domain. We have seen that the the Fourier transform of a normal multidimensional random vector, produces a set of real and imaginary coefficients that are also normal. Since the wavelet transform is a linear and orthogonal operation, the wavelet transform of the Fourier transform of a normal multidimensional random vector produces a normal complex vector. Let $f(t) \in L^2(\mathbf{R})$ and define its Fourier transform by

$$(\mathcal{F}f)(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt. \tag{4.50}$$

The continuous wavelet transform with scale a and shift b of the Fourier transform of f, is defined as

$$\mathcal{W}^{a,b}[(\mathcal{F}f)(\omega)] = \int_{-\infty}^{+\infty} (\mathcal{F}f)(\omega)\psi^*_{a,b}(\omega)d\omega, \tag{4.51}$$

where $(a,b) \in \mathbf{R}^+ \times \mathbf{R}$ and

$$\psi_{a,b}(\omega) = a^{-1/2}\psi\left(\frac{\omega - b}{a}\right). \tag{4.52}$$

Substituting equation 4.50 into equation 4.51, we get

$$\begin{aligned}
\mathcal{W}^{a,b}[(\mathcal{F}f)(\omega)] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(t)e^{-i\omega t}\,dt\psi^*_{a,b}(\omega)d\omega \\
&= \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-i\omega t}\psi^*_{a,b}(\omega)d\omega\,dt \\
&= \int_{-\infty}^{+\infty} f(t)\widehat{\psi^*_{a,b}}(t)\,dt,
\end{aligned} \tag{4.53}$$

where for real wavelets,

$$\widehat{\psi^*_{a,b}}(t) = \sqrt{a}e^{-i2\pi tb}\tilde{\psi}(at), \tag{4.54}$$

and $\tilde{\psi}$ is the Fourier transform of the mother wavelet ψ. Equation 4.53 represents the inner product of $f(t)$ with respect to the wavelet based function $\widehat{\psi^*_{a,b}}(t)$. In other words, $\mathcal{W}^{a,b}[(\mathcal{F}f)(\omega)]$ represents the similarity between $f(t)$ and the function $\widehat{\psi^*_{a,b}}(t)$, which acts like a window on the signal, $f(t)$.

By applying the SURE thresholding techniques to the real and imaginary wavelet-Fourier details (see figure 4.14), we can eliminate most of the noise at each decomposition level.

A variation of this purely wavelet-Fourier domain scheme may be employed to minimize the phase distortion introduced by the nonlinear effect of the thresholding techniques (STT and HTT)

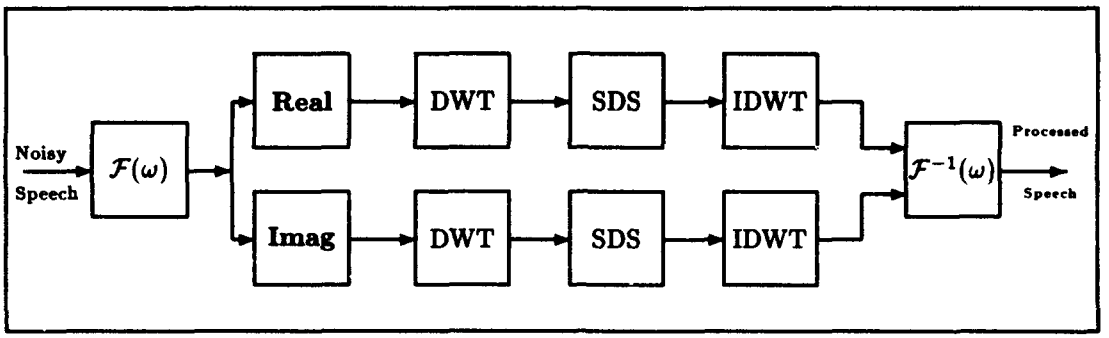

Figure 4.14 Speech de-noising in the frequency domain using wavelets

on the real and imaginary parts of the wavelet-Fourier details. In order to reduce the effect of phase distortions, we may save the noisy phase from the Fourier transform of the noisy voiced speech and restore it after the de-noising procedures. Figure 4.15 illustrates the method; the time domain voiced speech waveform is first Fourier transformed to extract the phase and then the wavelet transform of the Fourier transform is taken before the independent de-noising process of the real and imaginary parts is applied. The thresholded details (real and imaginary), are then inverse wavelet transformed independently in order to produce the de-noised real and imaginary parts, Fourier transformed in order to extract the de-noised amplitude, and finally, the old phase is combined with this newly calculated amplitude and inverse Fourier transformed back to the time domain. Observe, this method requires two Fourier transforms and four wavelet transforms (see figure 4.15).

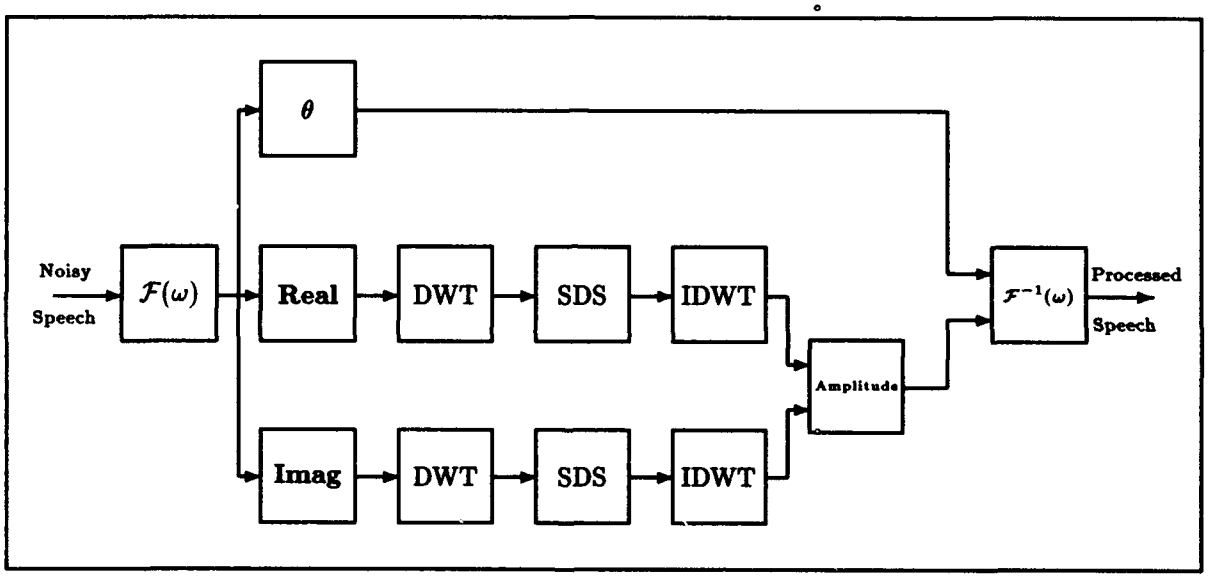

Figure 4.15 Speech de-noising in the frequency domain using noisy phase and wavelets

V. Experiments And Results

5.1 Experiments

In this chapter, we present the results of applying the thresholding techniques we developed in the last chapter. Eight different speech processing systems were studied. We start by explaining the assumptions made and the parameters used for each experiment. We then discuss the quantitative and qualitative results for all eight experiments, as well as the spectrum analysis for some experiments. The qualitative results are based on the total squared error between each experiment's output and both the clean and noisy signals. On the other hand, the qualitative results are all based on the results of the listening tests that we conducted with an untrained jury of six students (four males and two females). Before each informal listening test, the listener is given a chance to listen to both the clean and noisy speech speech signals (SNRs of 0db and 6db) and then he or she is briefed about what the test is all about (see figure 5.1). The listeners were asked to make a choice between two de-noised speech signals (e.g., choice between time processing vs. Fourier processing of the same noisy signal). Finally, we present and analyze some spectrograms of four different de-noising methods. We conclude this chapter with a summary of the tests' results and some of the recommendations we encountered throughout this thesis work.

5.1.1 Experimental Set Up. Due to the large number of methods and the flexibility of the parameters available for experiments, we fixed the following inputs to the speech de-noising algorithms we presented in chapter four:

1. The percent factor p applied to the unvoiced and silent portions is $p = 50\%$.
2. The maximum voiced decomposition level is $m_v = 6$.
3. The overlap between adjacent speech windows is $overlap = 16$.
4. The number of samples of the original speech ("They enjoy it when I audition") is $N = 31200$.
5. The sampling frequency is 16kHz.
6. The approximation coefficients (pitch of voiced speech) are not processed (i.e., untouched and still noisy).

Experimentally, we fixed the overlap between adjacent windows ($p = 16$), and we determined that by keeping only $p = 50\%$ of the ratio R_v (see equation 4.4), the transition obtained between the voiced portions to both the silent and the unvoiced portions improved intelligibility considerably.

5.1.2 Experimental Speech Signals. Starting with a clean speech signal ("They enjoy it when I audition") of 31200 samples, we generated seven different white Gaussian noise signals and seven noisy signals such that the signal-to-noise-ratios (SNRs) are as follows: -10db, -6db, -3db, 0db, 3db, 6db, and 10db. Using these noisy signals, we produced both soft thresholded and hard thresholded signals with the following methods:

1. De-noising in the time domain.
2. De-noising in the time domain using the noisy phase.
3. De-noising in the frequency domain.
4. De-noising in the frequency domain using the noisy phase.
5. De-noising in the time domain using wavelets.
6. De-noising in the time domain using the noisy phase and wavelets.

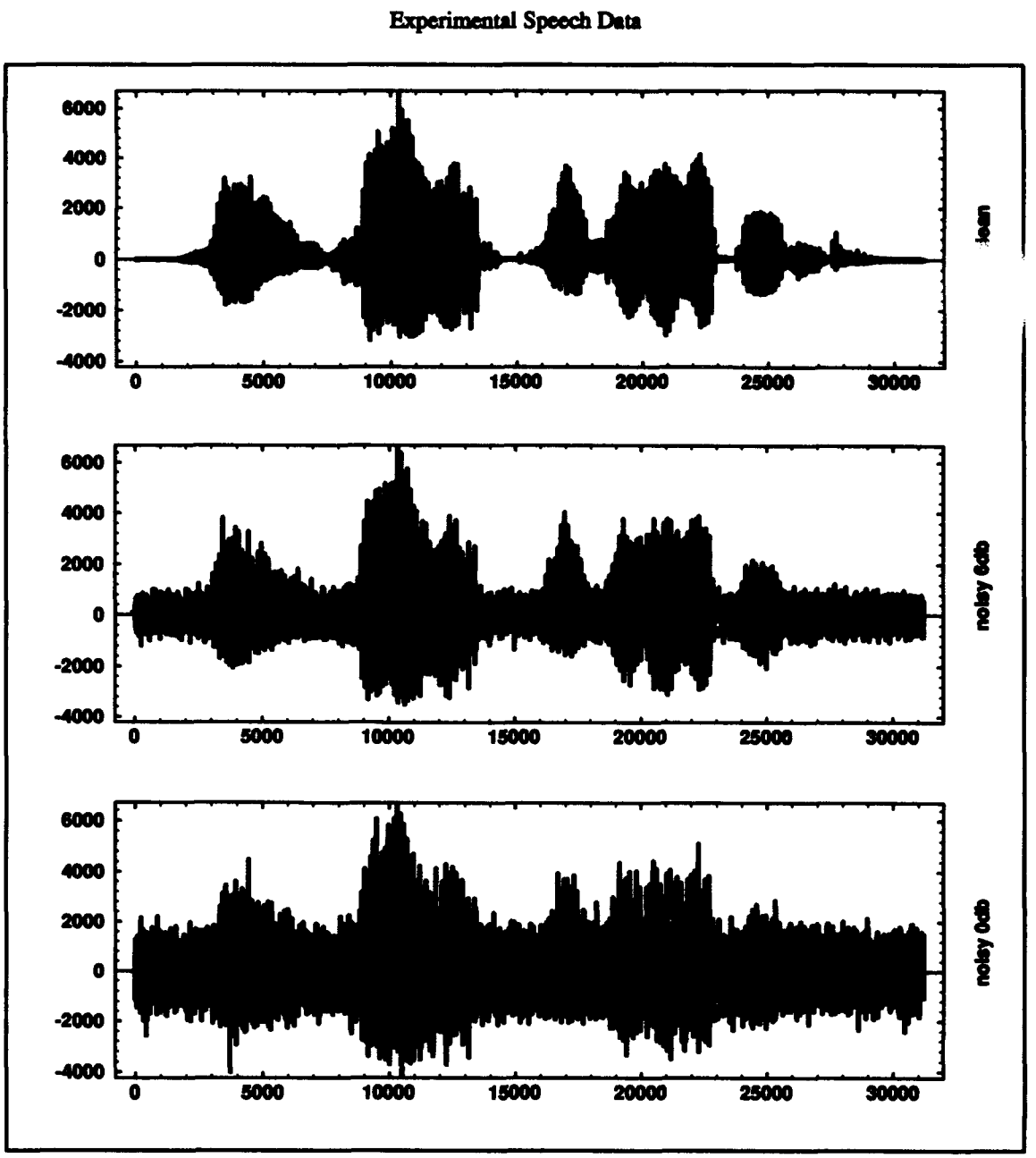

Figure 5.1 Clean speech and noisy speech (6db and 0db SNRs).

7. De-noising in the frequency domain using wavelets.

8. De-noising in the frequency domain using the noisy phase and wavelets.

The total number of de-noised signals without wavelets (i.e., using Steiņ's criteria) is 56. Since there are two thresholding techniques (STT and HTT), seven different noisy signals, and four different methods that don't involve wavelets. The total number of de-noised signals with wavelets (i.e., using Donoho's criteria) is 168, using two thresholding techniques (STT and HTT), seven different noisy signals, four different methods that involve wavelets, and three different wavelets used (db20, db6, and coiflets(6)). Hence, the total number of files studied is 224.

5.1.3 Quantitative analysis. Since the soft thresholding technique (STT) pulls towards zero *every* single voiced sample and the hard thresholding technique (HTT) pulls towards zero *only* the voiced elements below the hard threshold (t^{hard}) in absolute value, *theoretically*, we expect that the energy of the de-noised signal under the STT to be less than the energy of the de-noised signal under the HTT. In order to quantify this result, the total squared error between the de-noised signal and the noisy signal, using the STT technique, is defined as

$$Error_{STT}^{noisy} = \sum_{n=0}^{N-1} \left(x_n^{noisy} - x_n^{STT}\right)^2, \tag{5.1}$$

where N is the total number of samples of the speech signal under analysis (i.e., $N = 31200$). Similarly, the HTT total error is defined as

$$Error_{HTT}^{noisy} = \sum_{n=0}^{N-1} \left(x_n^{noisy} - x_n^{HTT}\right)^2. \tag{5.2}$$

Both $Error_{STT}^{noisy}$ and $Error_{HTT}^{noisy}$ measure the closeness of the de-noised signal to the noisy signal. Ideally, we want to be as far away as possible from the noisy signal, and still preserve the intelligibility of the de-noised speech signal. The experiments illustrate that (see figures F.1 and

F.3)

$$Error_{STT}^{noisy} \geq Error_{HTT}^{noisy}. \tag{5.3}$$

In fact, because of the definitions of the STT and the HTT, the use of the STT removes noise from *all* samples, while the use of the HTT removes noise *only* from certain samples. For this reason, the de-noised speech signal under the HTT has more remaining noise than the de-noised speech signal under the STT. In all experiments and for all seven noisy speech signals analyzed, figures F.1 and F.3, illustrate the fact that the STT outperforms the HTT with respect to the total squared error between the de-noised signals and the noisy signals.

Since the purpose of our de-noising technique is to attenuate the effect of the noise, we would like the de-noised speech signals to be as close to the clean signal as possible. In order to quantify this result, the total squared error between the de-noised signal and the clean signal, using the STT technique, is defined as

$$Error_{STT}^{clean} = \sum_{n=0}^{N-1} \left(x_n^{clean} - x_n^{STT}\right)^2. \tag{5.4}$$

Similarly, the HTT total error is defined as

$$Error_{HTT}^{clean} = \sum_{n=0}^{N-1} \left(x_n^{clean} - x_n^{HTT}\right)^2. \tag{5.5}$$

Both $Error_{STT}^{clean}$ and $Error_{HTT}^{clean}$ measure the closeness of the de-noised signal to the clean signal. Ideally, we want the de-noised speech signal to be as close as possible to the clean signal, and still preserve the intelligibility of the de-noised speech signal. Both theory and experiments

prove that (see figures F.2 and F.4)

$$Error_{STT}^{clean} \leq Error_{HTT}^{clean}. \tag{5.6}$$

Again, in all experiments and for all seven noisy speech signals analyzed, figures F.2 and F.4, illustrate the fact that the STT outperforms the HTT with respect to the total squared error between the de-noised signals and the clean signal.

5.1.4 Qualitative Analysis Of The Informal Listening Tests. The qualitative analysis of the de-noised speech signals, depends on many factors. In order to understand the advantages and disadvantages of each of the eight methods, described earlier, we chose to study two noisy signals with signal to noise ratios 0db and 6db, respectively. The 0db signal represents a noisy speech signal with a relatively high level of noise, while the 6db signal represents a noisy speech signal with a relatively low level of noise.

5.1.4.1 Effects Of STT vs. HTT. In order to study the effects of the STT vs. the HTT, we randomly selected a jury of six students (four males and four females), considered to be untrained listeners. We presented to these listeners two groups of speech signals; group **A** has speech signals processed using the STT method and group **B** has speech signals processed using the HTT method. Each group has two sets of de-noised speech signals, where the original noisy speech signals have SNRs of 0db and 6db. Each set has speech data processed using the following speech de-noising systems (SDS):

1. De-noising in the time domain.
2. De-noising in the time domain using the noisy phase.
3. De-noising in the frequency domain.
4. De-noising in the frequency domain using the noisy phase.
5. De-noising in the time domain using wavelets.

6. De-noising in the time domain using the noisy phase and wavelets.

7. De-noising in the frequency domain using wavelets.

8. De-noising in the frequency domain using the noisy phase and wavelets.

We asked the students to listen to each speech signal from group **A** and compare it with its counterpart in group **B** (e.g., 0db of group **A** using SDS-1 vs. 0db of group **B** using SDS-1). All the students, concluded that the STT method has less remaining noise and, hence, it is easier to listen to the STT-processed speech signals than the HTT-processed speech signals. For this reason, we chose to continue experimenting with only the speech signals produced by the STT method.

5.1.4.2 Effects Of Preserving The Noisy Phase Using STT. Based on the results of the STT vs. the HTT experiment, above, and in order to study the effects of the phase, we presented to the same jury of students, two groups of speech signals processed using the STT method; group **A** has de-noised speech data processed without restoration of the noisy phase and group **B** has de-noised speech data with restoration of the noisy phase. Each group has two sets of de-noised speech signals, where the original noisy speech signals have SNRs of 0db and 6db. Each set has speech data processed using the following speech de-noising systems (SDS):

A. No preservation of the noisy phase:

1. De-noising in the time domain.

2. De-noising in the frequency domain.

3. De-noising in the time domain using wavelets.

4. De-noising in the frequency domain using wavelets.

B. Preservation of the noisy phase:

1. De-noising in the time domain using the noisy phase.

2. De-noising in the frequency domain using the noisy phase.

3. De-noising in the time domain using the noisy phase and wavelets.

4. De-noising in the frequency domain using the noisy phase and wavelets.

We asked the students to listen to each speech signal from group **A** and compare it with its counterpart in group **B** (e.g., 0db of group **A** using SDS-1 vs. 0db of group **B** using SDS-1). All the students, concluded that intelligibility of group **B** is much better than the intelegibility of **A** and it is easier to listen to the speech signals processed with noisy phase restoration than to listen to the speech signals processed without noisy phase restoration. For this reason, we chose to continue experimenting with only the speech signals produced using both the STT method and the phase restoration technique.

5.1.4.3 Effects Of The Time vs. Fourier Domains On Speech De-noising Using STT And Noisy Phase Restoration. Based on the results of the last two sections, we chose to continue experimenting with speech data processed using both the STT and the phase restoration techniques. In order to study the effect of the time domain vs. the Fourier domain, we presented to the jury of students, two groups of de-noised speech signals; group **A** has speech data processed in the time domain and group **B** has speech data processed in the Fourier domain. Both groups have speech data processed using both the STT and phase restoration techniques. Each group has two sets of de-noised speech signals, where the original noisy speech signals have SNRs of 0db and 6db. Each set has speech data processed using the following speech de-noising systems (SDS):

A. Time domain:

1. De-noising in the time domain using the noisy phase.
2. De-noising in the time domain using the noisy phase and wavelets.

B. Fourier domain:

1. De-noising in the frequency domain using the noisy phase.
2. De-noising in the frequency domain using the noisy phase and wavelets.

We asked the students to listen to each speech signal from group **A** and compare it with its counterpart in group **B** (e.g., 0db of group **A** using SDS-1 vs. 0db of group **B** using SDS-1). All the students, concluded that the intelligibility of group **B** is much better than the intelligibility of **A**

and it is easier to listen to the speech signals processed in the Fourier domain than to listen to the speech signals processed in the time domain. For this reason, we chose to continue experimenting with only the speech signals produced in the Fourier domain using both the STT method and the phase restoration technique.

5.1.4.4 Effects Of Wavelets On Speech De-noising In The Fourier Domain. Based on the results of the last three sections, we chose to continue experimenting with speech data processed in the Fourier domain using both the SST and the phase restoration techniques. In order to study the effect of using wavelets vs. not using wavelets in the Fourier domain, we presented to the jury of students, two groups of de-noised speech signals; group **A** has speech data processed in the Fourier domain without using wavelets and group **B** has speech data processed in the Fourier domain using wavelets. Both groups have speech data processed in the Fourier domain using the STT and phase restoration techniques. Each group has two sets of de-noised speech signals, with SNRs of 0db and 6db. Each set has speech data processed using the following speech de-noising systems (SDS):

A. wavelets:

1. De-noising in the frequency domain using the noisy phase and wavelets.

B. No wavelets:

1. De-noising in the frequency domain using the noisy phase.

We asked the students to listen to each speech signal from group **A** and compare it with its counterpart in group **B** (e.g., 0db of group **A** using SDS-1 vs. 0db of group **B** using SDS-1). All the students, concluded that for 6db, the de-noised speech signals from both groups are very close in terms of intelligibility, however, for 0db, the intelligibility of group **A** is much better than the intelligibility of **B**. Since our jury is forced to choose only one group, all students chose group **A** because they concluded that it is easier to listen to the speech signals processed in the Fourier

domain using wavelets than to listen to the speech signals processed in the Fourier domain without using wavelets.

5.1.5 Spectrum Analysis Of De-noised Speech Data Using The STT. We mentioned earlier that the production of speech through the vocal tract is characterized as either voiced or unvoiced. The unvoiced speech signals, the fricatives, behave like noise and have high energy above about 3kHz and relatively very low energy below 3kHz (19). On the other hand, most voiced speech is located at bands of frequencies below 3kHz. The pitch and the first formant are, in general, located below $500Hz$, while the second and third formants are located between $500Hz$ and 3kHz. The formant frequencies are important because most of the voiced speech characteristics (i.e., pitch) are based on the location of these frequencies. In order to study the effects on the frequency content of our speech signals, we choose three different wavelets and four different de-noising techniques. We generated two sets of spectrograms, wide-band and narrow-band (for clean and noisy speech, only). Wide-band spectrograms have a small analysis window, therefore, the frequency resolution is low, while the time resolution is high. On the other hand, narrow-band spectrograms have a large analysis window, therefore, the frequency resolution is high, while the time resolution is low. Each set (narrow-band and wide-band) of spectrograms includes :

1. Clean speech.

2. Noisy speech 0db (relatively high noise level).

3. Noisy speech 6db (relatively low noise level).

4. for each of the three wavelets used (db20, db6, and coiflets(6)) and for each of the noise levels used (0db and 6db), we studied the frequency content of the de-noised speech data using shrinkage with the following speech de-noising systems (SDS):

   a. De-noising in the time domain.

   b. De-noising in the time domain using wavelets.

c. De-noising in the frequency domain using the noisy phase.

d. De-noising in the frequency domain using the noisy phase and wavelets.

The spectrograms of the clean signal show very clearly the pitch, the first, second, and third formants. These frequencies have high energy below 3kHz (see figure G.1). Despite the addition of the white Gaussian noise, the spectrograms of the noisy speech signals with signal-to-noise ratios of 0db and 6db, show that the pitch, the first, second, and third formants are still dominant below the 3kHz frequency. However, the effect of white Gaussian noise can be clearly seen throughout the spectrograms. In fact, since the white Gaussian noise is, in general, a broad-band signal, the spectrogram indicates high energy at all frequency bands (see figures G.2 and G.3).

5.1.5.1 Effects Of Stein's Criteria On Time De-noising vs. Fourier De-noising Using Noisy Phase Restoration. De-noising in the time domain using both Stein's criteria and the noisy phase, works relatively well for high signal-to-noise ratios. In fact when the noise level is very low (i.e., 6db), most of the signal's formant's structure below the 3kHz frequency is still preserved; however, a lot of high frequency noise is still present (see figure H.1). When the noise level increases, the noisy speech signal looks like white Gaussian noise and the application of Stein's criteria tends to eliminate most of the speech signal itself, and hence affecting most of the formant frequencies. On the other hand, de-noising in the Fourier domain using Stein's criteria and preserving the noisy phase, works much better because of the fact that the noise is split between the real and imaginary parts of the Fourier transform. Since the noisy phase is restored, most of the noisy speech structure (pitch and formants) is restored back to the de-noised speech signal. Despite these improvements, when the noise level is relatively high, the real and imaginary parts become very noisy and Stein's criteria affects the true structure of the signal (see figure H.2).

5.1.5.2 Effects of The wavelet Choice On De-noising In The Time Domain. By using wavelets, we decompose a noisy signal into bands of frequencies and then we de-noise each band separately. This process is potentially more powerful than the methods that don't use wavelets.

However, the choice of the right wavelet with good filtering characteristics is very important. We choose three different wavelets: db6, coiflets(6), and db20. Since the wavelet transform is a filtering operation, the effect of the filtering characteristics of the wavelet become very crucial. The spectrograms for both 0db and 6db using wavelets in the time domain show that there is an aliasing effect for both db6 and coiflets(6) (see figures I.1 and I.2). The reason for this aliasing is due to the fact that the Fourier transforms of both db6 and coiflets(6) have a lot of high energy side lobes which cause the filtering qualities of these wavelets to be of low importance. On the other hand, the spectrograms of db20 show no aliasing at all, which make db20 a very good wavelet to use in speech processing (see figure I.3). However, because of the fact that the cubic splines are not compactly supported wavelets, their use in practice requires an approximations which affects the general behavior of the spline wavelets. The best results in terms of total square error, intelligibility, and the preservation of formant frequencies , were given by db20 which is a compactly supported wavelet with a very good filtering quality (i.e., very small side lobes).

5.1.5.3 Effects Of The Wavelet Choice On De-noising In The Fourier Domain With Noisy Phase Restorations. Since the de-noising process is carried out in the Fourier domain, the noise level is split between the real and imaginary parts. These are then wavelet transformed and decomposed into bands of frequencies in order to eliminate most of the noise from each band. By restoring the noisy phase and applying the wavelet shrinkage to both the real and imaginary parts of the Fourier transform of the noisy signal, the effect of aliasing seems to decrease, even for db6 and coiflets(6) (see figures J.1 and J.2). However, for the same reasons described in the previous section, most of the formants' structure of the noisy speech signal is preserved when using db20 (see figure J.3).

5.2 Conclusions

In this chapter, we presented the results of several speech de-noising experiments on various noisy speech data (-10db to 10db). In general, we saw that the performance of the speech de-noising systems using both Fourier and wavelets resulted in intelligible speech even for low signal-to-noise ratios (SNR). The use of the noisy phase improved both the quality and the intelligibility of the de-noised speech signals. The use of the soft thresholding technique (STT), in the wavelet-Fourier domain, proves to be a very good technique to use in the enhancement of noisy speech data.

VI. Conclusions and Recommendations

6.1 Introduction

In this chapter, we present both the conclusions of this research and some of the recommendations for future research in the area of enhancing noisy speech data. We summarize the major points and evaluate how well the objectives of this thesis were met.

6.2 Main Conclusions Of The Thesis

This thesis is successful in producing several speech de-noising systems (SDS) in the time, Fourier, and the wavelet domains. Without the use of wavelets, the SDS systems perform relatively well and produce intelligible speech when the noise level is low (SNR = 6db). These systems are comparatively fast (since they don't require the wavelet transform) and can be used to produce comparable results to the wavelet-based SDSs, for low levels of noise (e.g., SNR = 6db). However, when the noise level is high (SNR = 0db), the non-wavelet SDSs do not produce intelligible speech. In fact, without using wavelets, the application of either the soft thresholding technique (STT) or the hard thresholding technique (HTT) to noisy speech data, with noise levels below SNR = 6db, produced de-noised speech data, that is worst to listen to, than the noisy speech data itself.

The application of Stein's criteria to noisy voiced speech using wavelets on the time data (Donoho's technique) did not produce intelligible speech for all noise levels (i.e., -10db to 10db). In fact, this method produced a very distorted de-noised speech with a constant disturbing sound, which is mainly due to the non-linear effect of the thresholding techniques. The use of the noisy phase produced a slight improvement of the intelligibility of speech. Finally, the use of the wavelet shrinkage techniques applied to the Fourier domain with noisy phase restoration proves to be a powerful technique to enhance speech data degraded by additive white Gaussian noise. In fact, when using a wavelet with good filtering characteristics (e.g., db20), the formants' structure and intelligibility can be considerably preserved. This new technique involves a lot of calculations

due to the Fourier transforms, the wavelet transforms, and the phase calculations. However, the intelligibility of the de-noised speech data, outperformed all the other de-noising systems, especially when the noise levels were high (SNRs below 6db).

The combination of the noisy phase and the wavelet-Fourier technique produced the best results (intelligibility) because it involves a de-noising process on two less noisy sets of data; the real and imaginary parts of the Fourier transform of the noisy signal. The Fourier transform splits the noise level between the real and imaginary parts. De-noising the wavelet details of both the real and imaginary parts, reduces the noise at each level of decomposition, resulting in a large amount of noise being taken from both the real and imaginary parts. After this de-noising process, the combination of the real and imaginary parts produces a cleaner amplitude which is further combined with the noisy phase, wherein important speech information is saved. Most importantly, this research illustrated the fact that the phase has the potential to preserve a lot of the underlying speech formants' structure and that, in order to avoid aliasing and still preserve intelligibility, it is very important to choose a wavelet with very good filtering characteristics.

6.3 Evaluation Of The Thesis Objectives

In terms of the four objectives mentioned in the first chapter, in this thesis, we were able to apply both wavelets and the soft thresholding technique (STT) to enhance noisy speech data. The speech de-noising systems (SDS) can only be applied to the voiced portions. The unvoiced and silent portions are not to be processed using the SDSs discussed in the fourth chapter. These portions tend to disappear when processed by the SDSs, and hence, we can use our SDSs as detector systems for the unvoiced, voiced, and silent speech portions by using a single window on the entire speech utterance. The use of the noisy phase, combined with both wavelets, Fourier, and the STT technique, considerably improved intelligibility. The use of wavelets with thresholding is important,

however, in order to obtain good results, the choice of a wavelet with good filtering characteristics (no high energy side-lobes) have a direct effect on the quality of the de-noised speech data.

6.4 Recommendations

Further investigations in the area of noise cancellation using both Fourier and wavelets can further the results of this research. Many of the methods described in this work can be further explored, improved (i.e., hard or soft thresholding of the approximations where the pitch of voiced speech resides), and compared to our results. The STT and HTT methods can be used to develop a pre-processing system to detect the voiced, unvoiced, and silent speech portions. Since this research assumes that the location of voiced, unvoiced, and silent portions are known, the STT or HTT based detector system, can complete our de-noising system.

One of the main concerns of our speech de-noising algorithm is speed. Due to the fact that our algorithm uses the Fourier transform, the wavelet transform, and the STT or the HTT techniques, the results tend to take considerable time to produce (an average of 8 minutes on a Sparc2 station with a single processor). In order to reduce the algorithm execution time, we suggest implementation of the algorithm in a parallel machine and we need also to derive a better way of finding the thresholds that minimize the SURE functions of either the HTT or the STT methods. In fact, the SURE function involves many loops and many comparisons that use each element of the noisy data. This means that as the number of data points increases, the execution time increases exponentially.

Finally, the results of this research illustrated the need for a better metric system for analyzing the performance of de-noising speech data. Most of the speech de-noising systems produced speech data with low L^2 error with respect to the clean speech signal, however, they do not have good intelligibility.

Most importantly, since we are using both wavelets and Fourier transforms, most of the processing can be implemented using parallel processing to speed up the results. Finally, the SURE functions should be further studied in order to find an effective criteria to choose the thresholds that minimize the SURE functions without checking all the samples available.

Appendix A. Wavelet Coefficients

This Appendix contains both the h filter coefficients and the Fourier transforms for each of the three wavelets used in this thesis: db6, coiflet(6), and db20. The Fourier transforms show clearly that the approximation filters, h, are low-pass filters, while the detail filters, g, are high-pass filters. These filters are used in the discrete wavelet transform (DWT) to divide the frequency spectrum, of the signal under analysis, into bands which have a constant bandwidth on a logarithmic frequency scale (in our case the bandwidths change by a factor of 2, the dilation factor). Observe that the Fourier transforms (g and h filters) of both db6 and coiflet(6) do not have a sharp roll-offs, while those of the Fourier transform of db20 are sharper than those of db6 and coiflets(6).

N	*n*	*coefficients of the filter h*
6	0	.332670552950
	1	.806891509331
	2	.459877502118
	3	-.135011020010
	4	-.085441273882
	5	.035226291882

Table A.1 Scaling function coifficients of db6.

N	*n*	*coefficients of the filter h*
6	0	-.07273261951
	1	.33789766250
	2	.85257202020
	3	.38486484700
	4	-.07273296500
	5	-.01565572800

Table A.2 Scaling function coifficients of coiflet(6).

N	*n*	*coefficients of the filter h*
20	0	.026670057901
	1	.188176800078
	2	.527201188932
	3	.688459039454
	4	.281172343661
	5	-.249846424327
	6	-.195946274377
	7	.127369340336
	8	.093057364604
	9	-.071394147166
	10	-.029457536822
	11	.033212674059
	12	.003606553567
	13	-.010733175483
	14	.001395351747
	15	.001992405295
	16	-.000685856695
	17	-.000116466855
	18	.000093588670
	19	-.000013264203

Table A.3 Scaling function coifficients of db20.

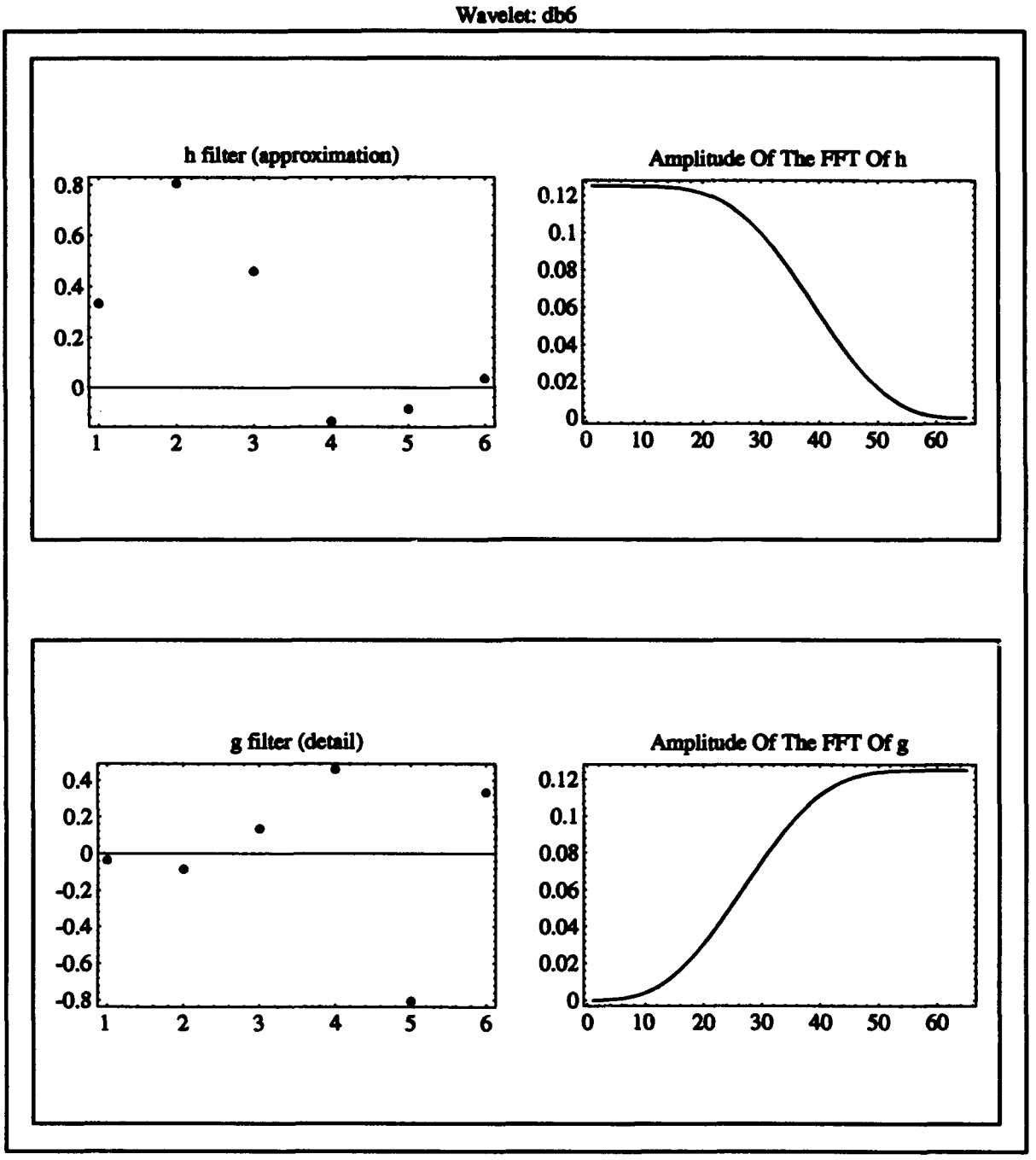

Figure A.1 Fourier transforms of the h and g filters of db6.

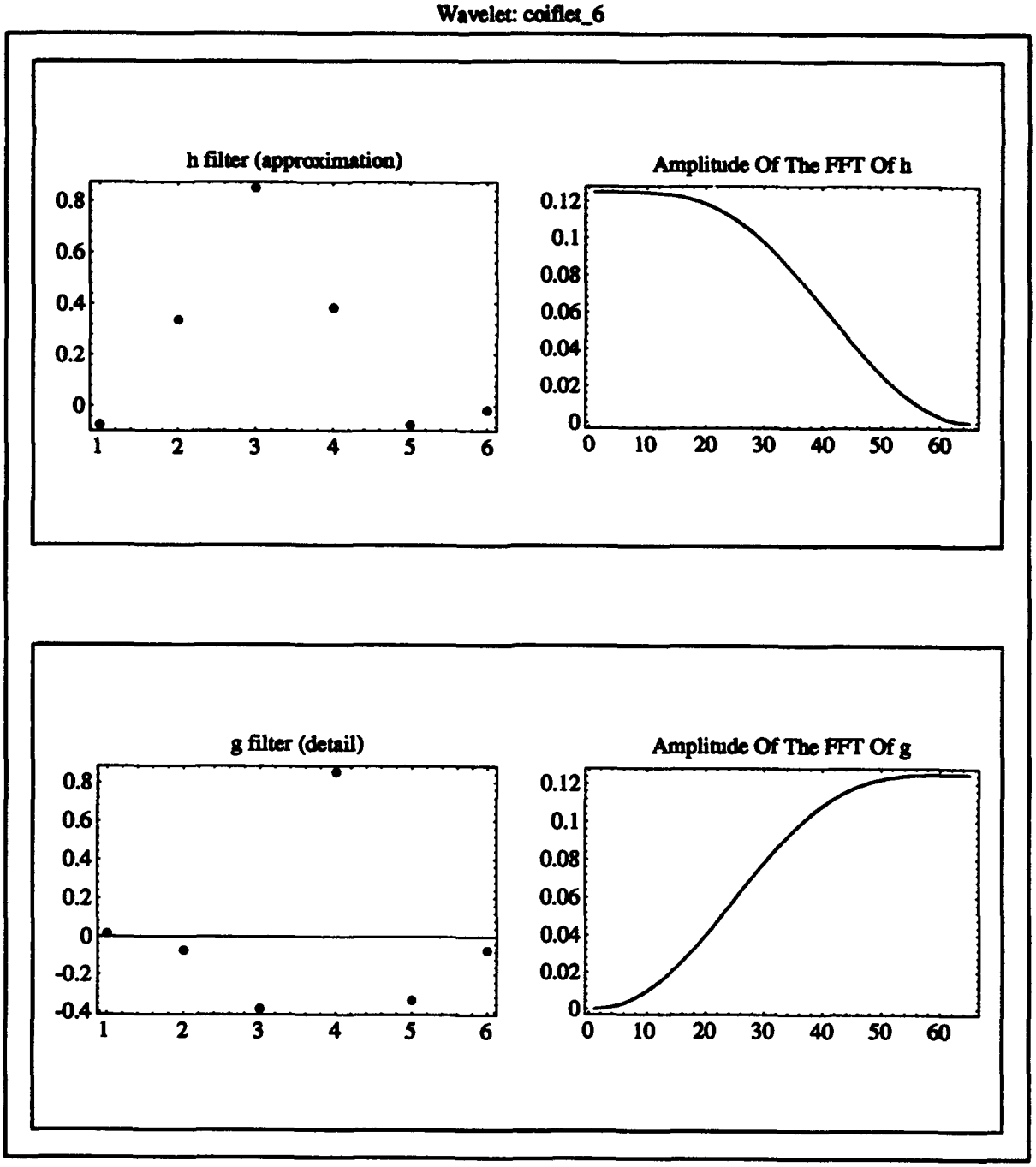

Figure A.2 Fourier transforms of the h and g filters of coiflet(6).

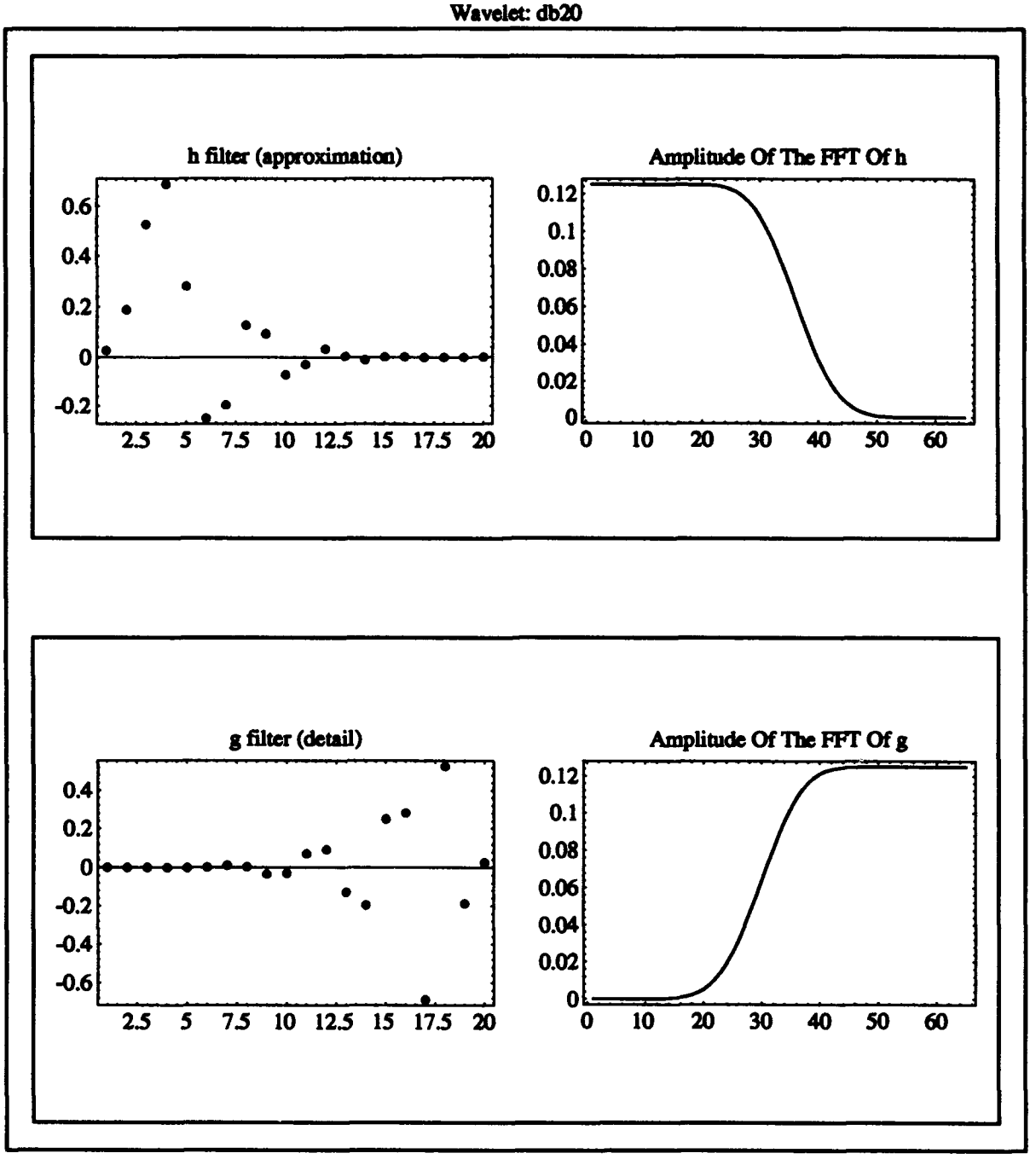

Figure A.3 Fourier transforms of the h and g filters of db20.

Appendix B. Wavelets And Their Fourier Transform

This Appendix contains the plots, on linear scales, of the three wavelets used in this thesis. All figures have identical time and frequency axes in arbitrary units. The amplitude of the Fourier transform of all three wavelets represent band-pass filters. Observe that the amplitudes of the Fourier transforms of db6 and coiflet(6), have many high energy side-lobes, while the amplitude of the Fourier transform of db20 has very little or no side-lobes at all. These filtering characteristics of the three wavelets affect the quality of the speech de-noising results (see spectrograms of Appendix J through L).

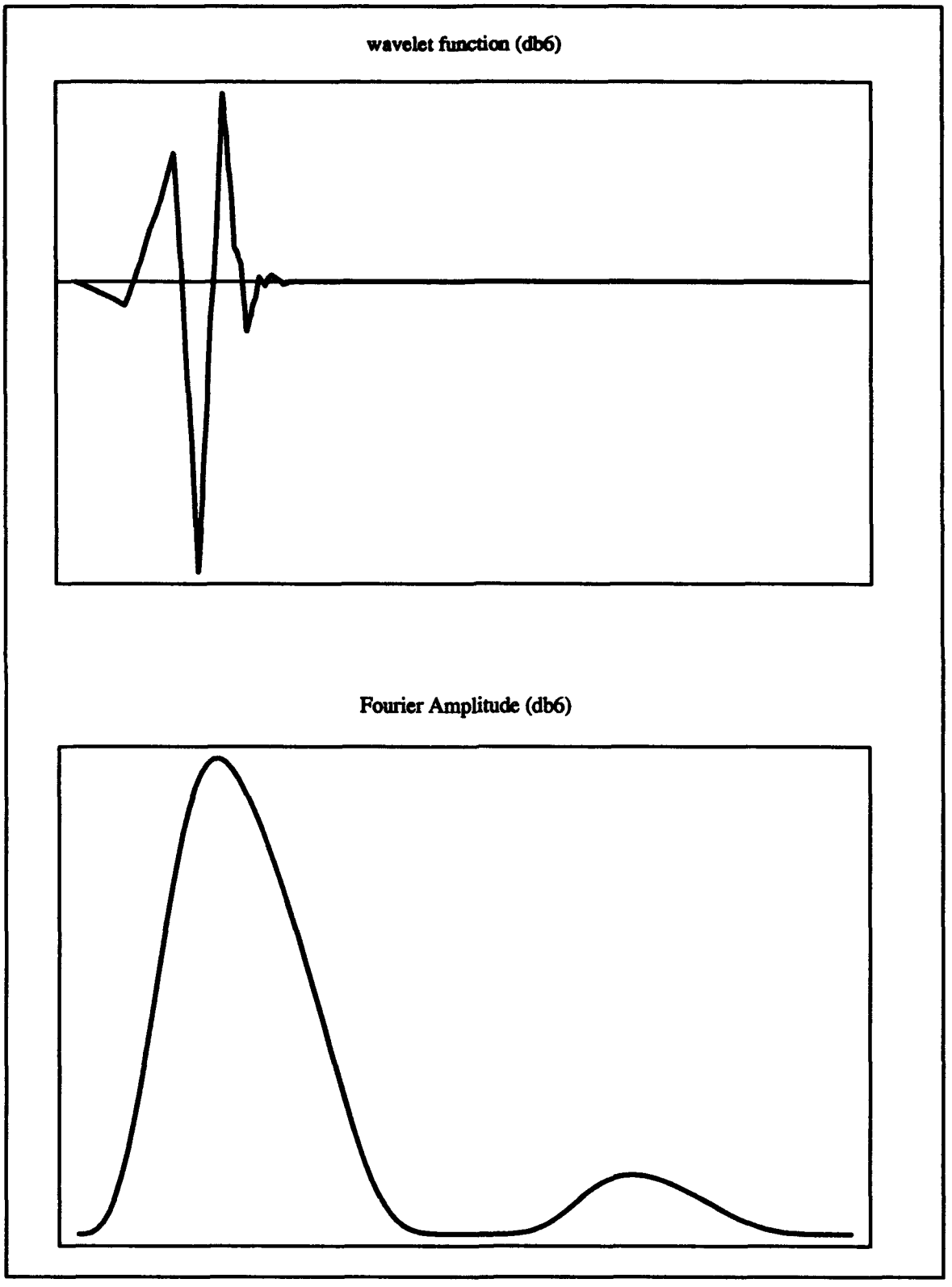

Figure B.1 Wavelet db6 and its Fourier transform.

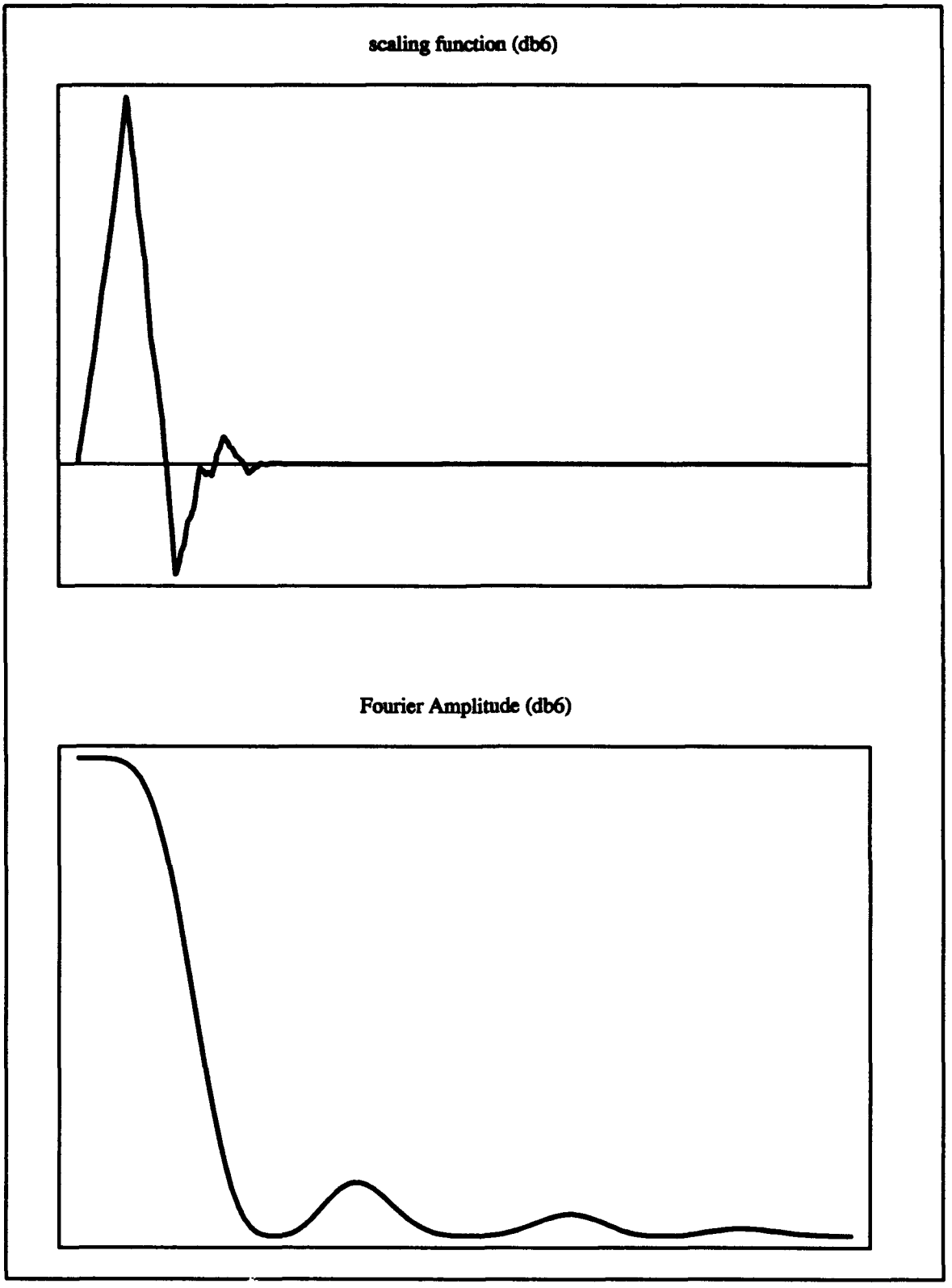

Figure B.2 Scaling function of the wavelet db6 and its Fourier transform.

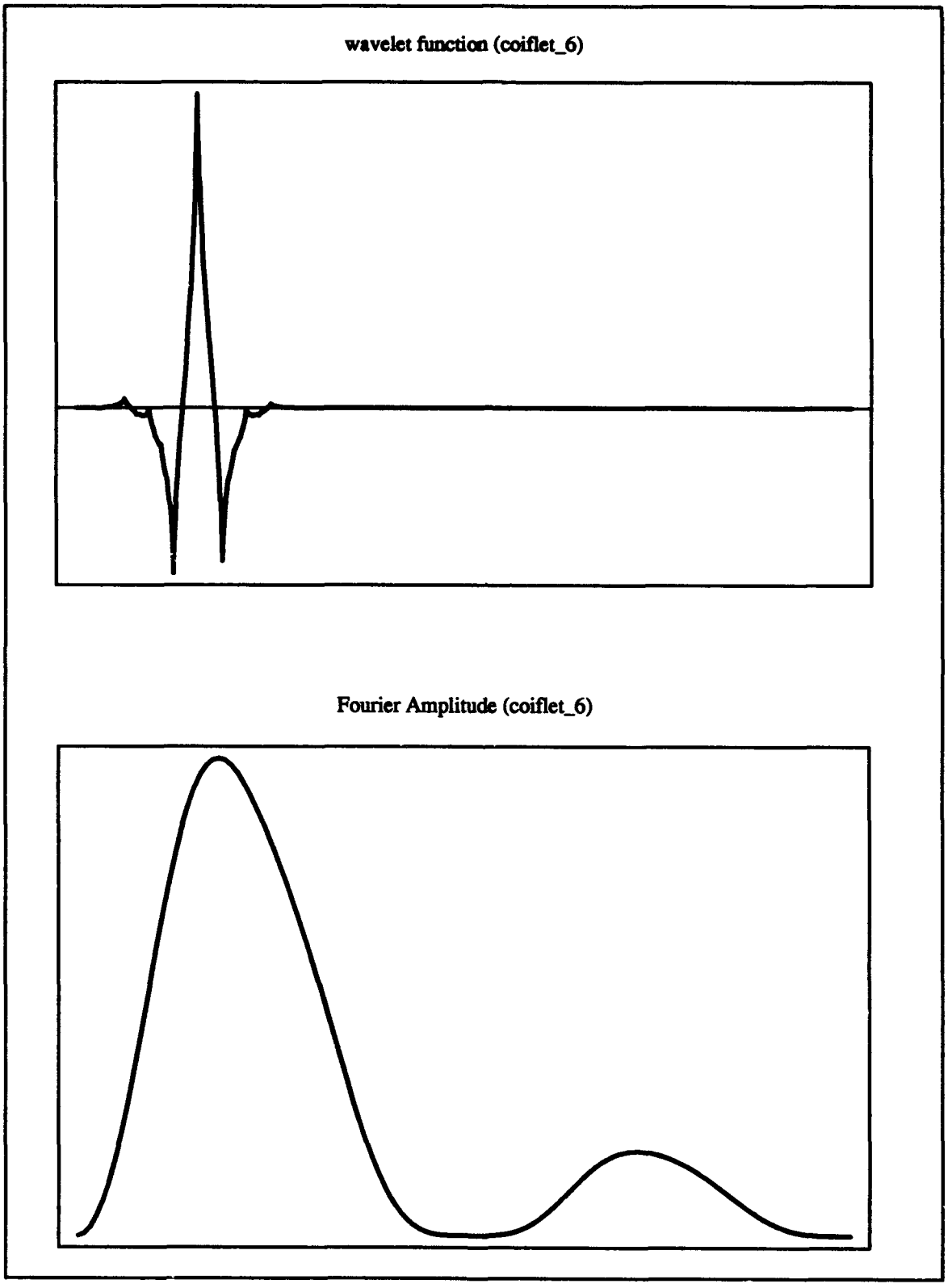

Figure B.3 Wavelet coiflet(6) and its Fourier transform.

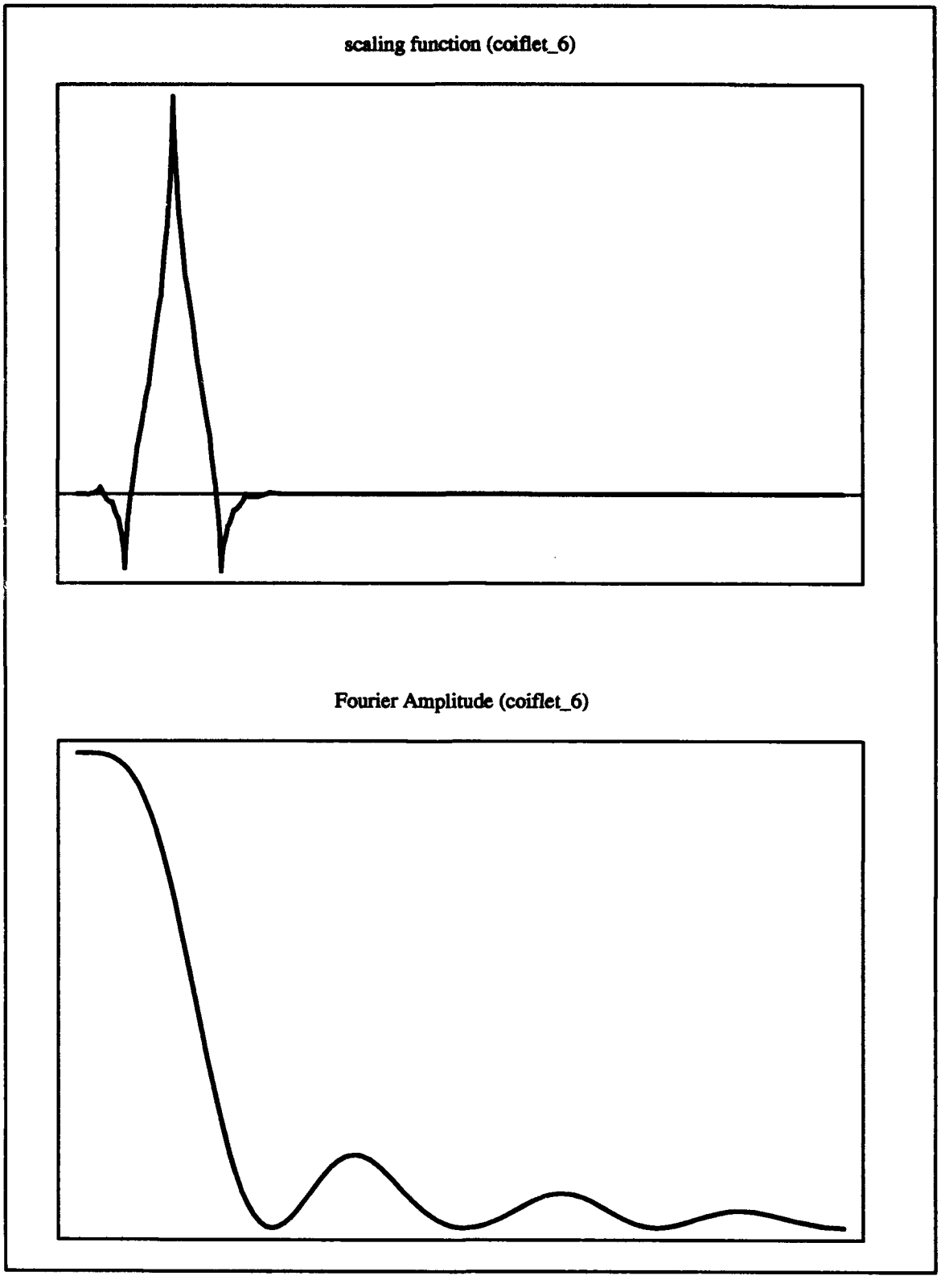

Figure B.4 Scaling function of the wavelet coiflet(6) and its Fourier transform.

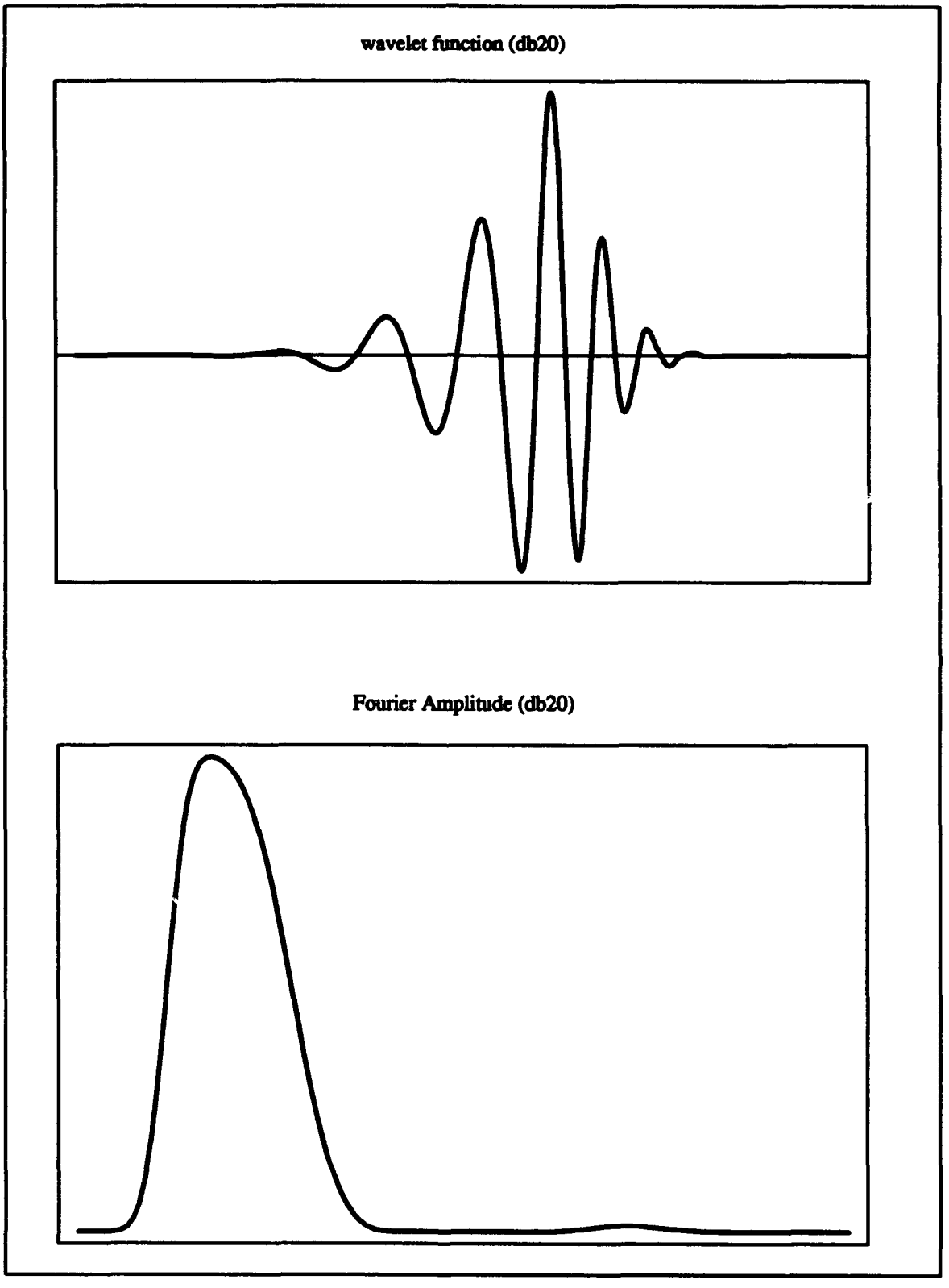

Figure B.5 Wavelet db20 and its Fourier transform.

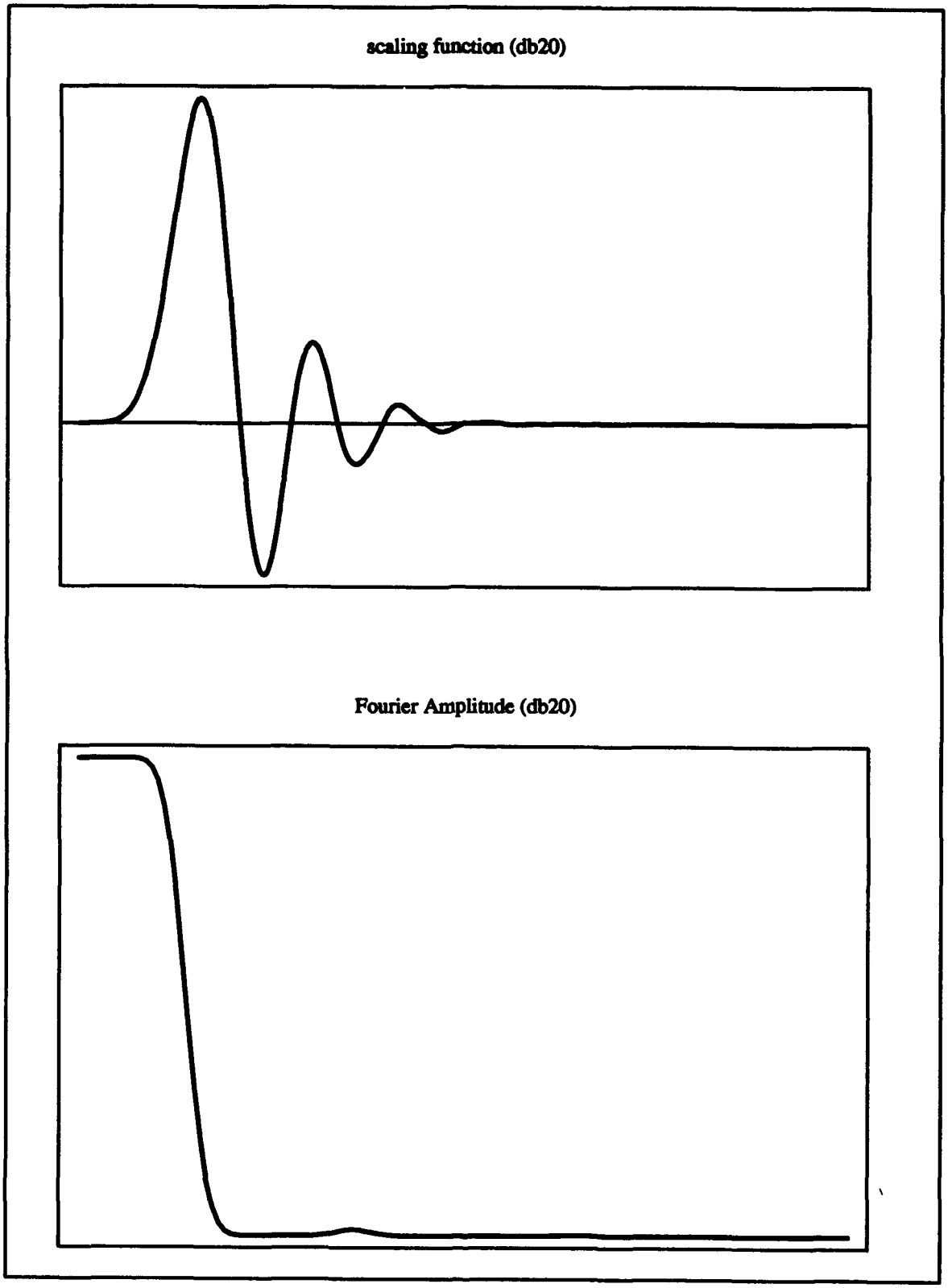

Figure B.6 Scaling function of the wavelet db20 and its Fourier transform.

Appendix C. Wavelet Shrinkage of Sinewave

This appendix contains the de-noising results of a sinewave of frequency 2Hz. We generated two signals, each contains 512 samples. The first signal is a 2Hz sinewave and the second is a white Gaussian noise of zero mean and variance of $\sigma^2 =$. We added the white Gaussian noise to the clean sinewave and then applied the soft thresholding technique (STT) to both the clean and noisy sinewaves. The de-noising process was carried in the wavelet domain using db20. The discrete wavelet transform (DWT) of the clean 2Hz sinewave, shows high energy details at the seventh and eighth levels, while the DWT of the noisy sinewave shows high energy details at all levels (see figures C.1 and C.2). The high energy details of the early levels of decomposition (i.e., levels 1, 2, 3, and 4) of the noisy sinewave, are mainly due to noise. We applied the STT to both the clean and noisy sinewaves, separately.

The clean signal was processed using the STT method and a variance value of $\sigma^2 = 1$. Figure C.3 shows that the details of the clean signal are still preserved and figure C.4 shows the near perfect reconstruction of the the clean sinewave. Observe, the amplitude of the Fourier transform of the STT processed clean signal is almost identical to that of the original clean sinewave (see figure C.5). Notice the phase distortions caused by the non-linear processing of this sinewave (see figure C.6).

The application of the STT to the noisy sinewave has eliminated most of the high frequency details which are mainly due to noise. Figure C.7 shows that the high energy details of the early levels of decomposition (i.e., levels 1, 2, 3, and 4) of the noisy sinewave, have been completely eliminated, while the details of the seventh and eighth levels of decomposition, which characterize the clean sinewave, are still preserved. The reconstructed sinewave, see figure C.8, is very close to the clean sinewave. Observe, the effects of the STT on both the amplitude and the phase of the Fourier transform of the noisy and reconstructed sinewaves (see figures C.9 and C.10).

Sinewave: Frequency = 2 And Variance = 1

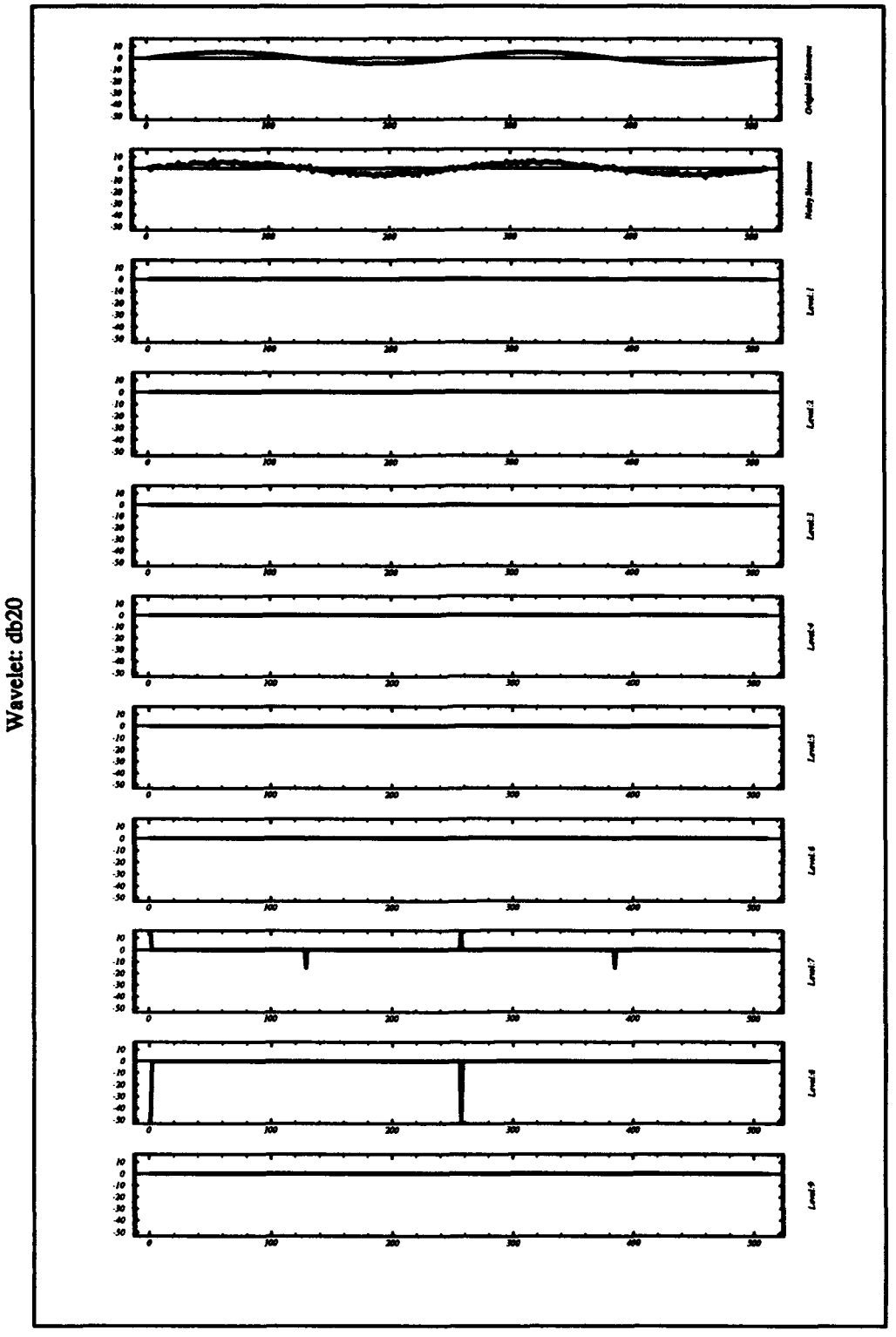

Figure C.1 Details of the clean sinewave (2Hz) .

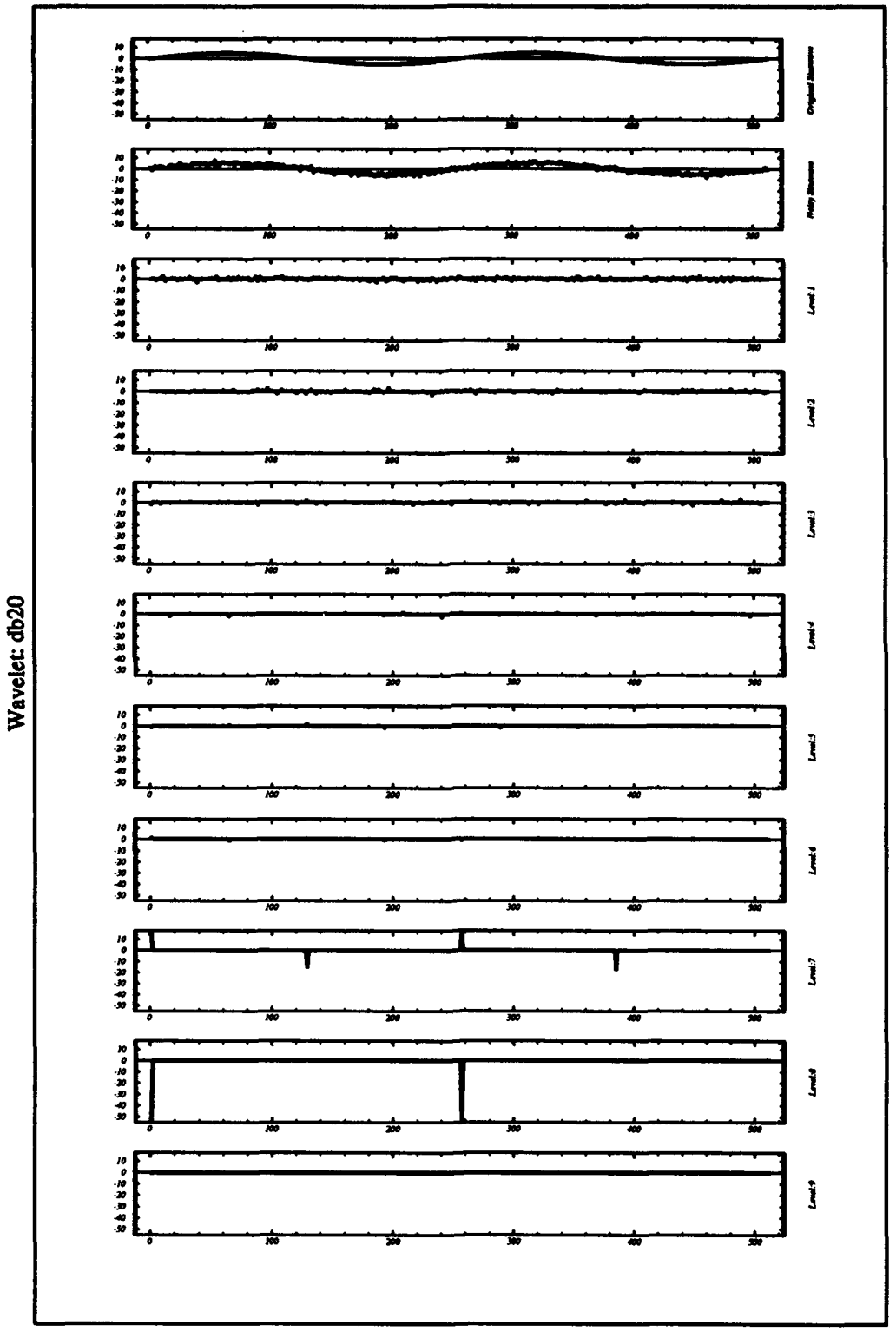

Figure C.2 Details of the noisy sinewave (2Hz).

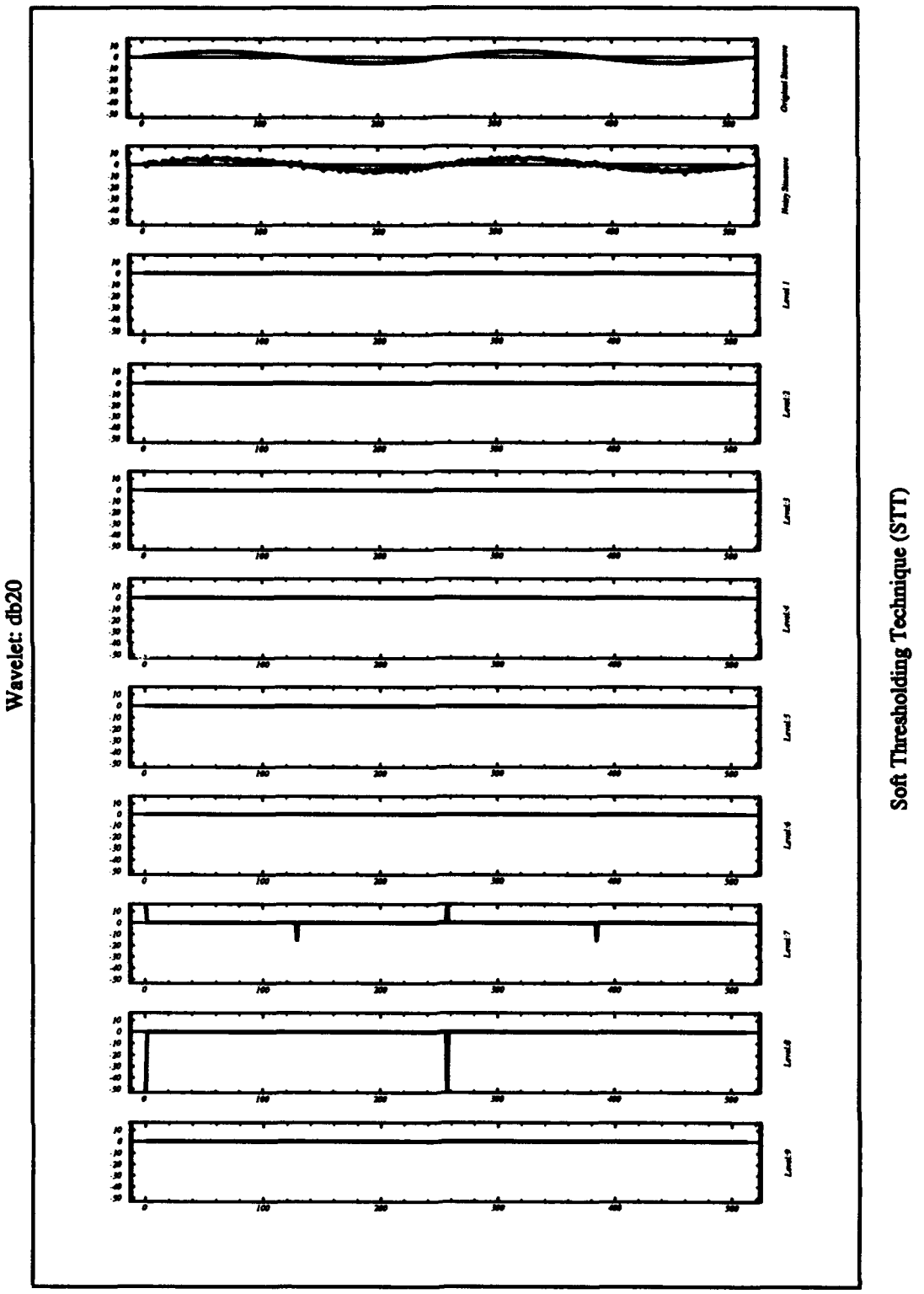

Figure C.3 Details of the processed clean sinewave (2Hz) after the STT ($\sigma^2 = 1$).

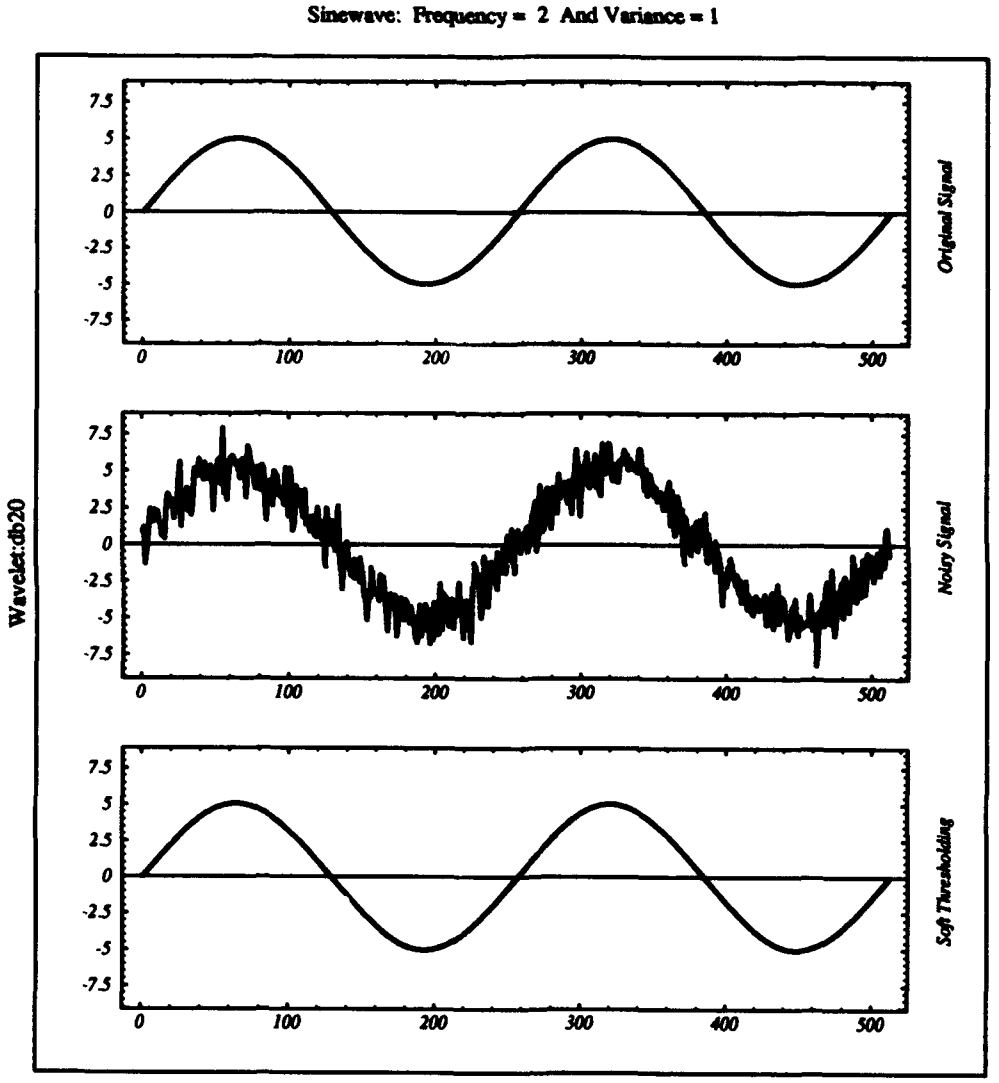

Figure C.4 Clean sinewave (2Hz) after the STT ($\sigma^2 = 1$).

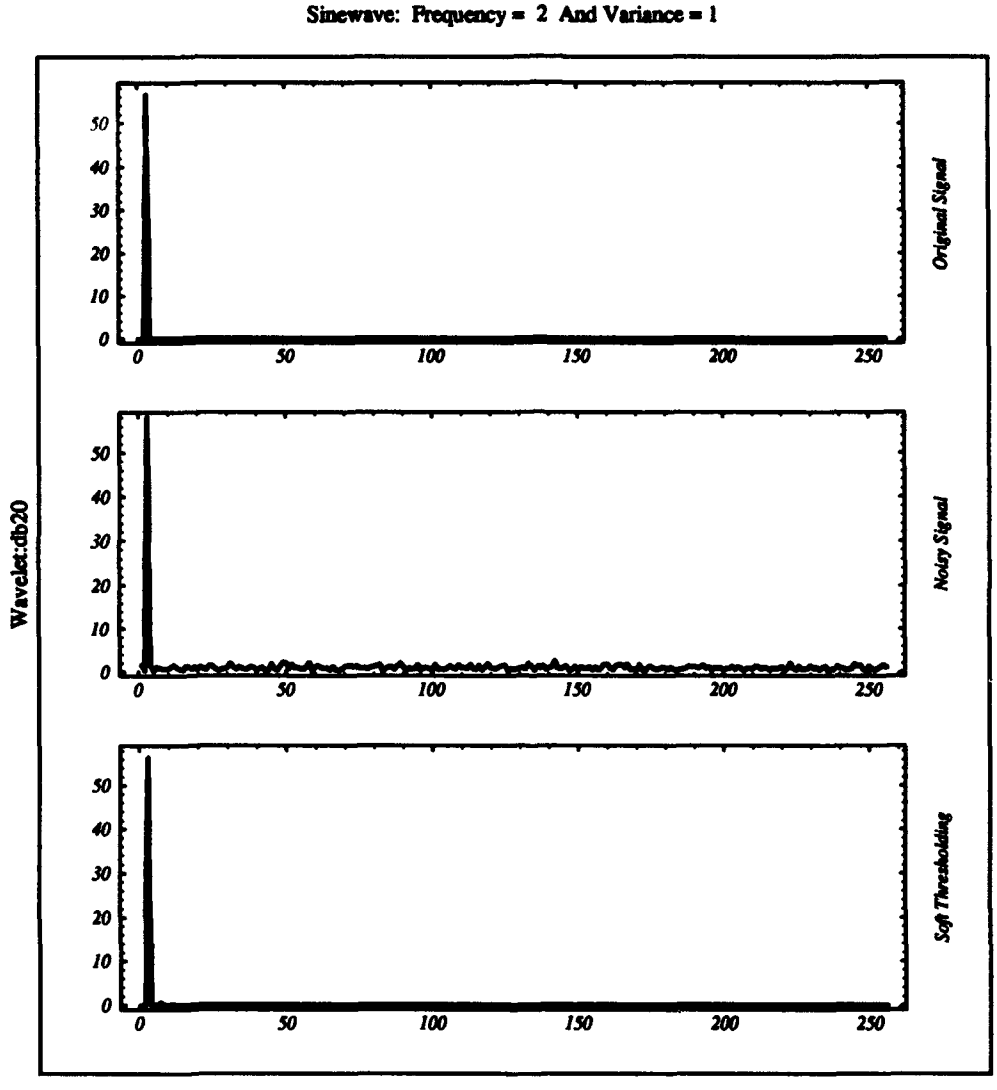

Figure C.5 Amplitude of the FFT of the clean sinewave (2Hz) after the STT ($\sigma^2 = 1$).

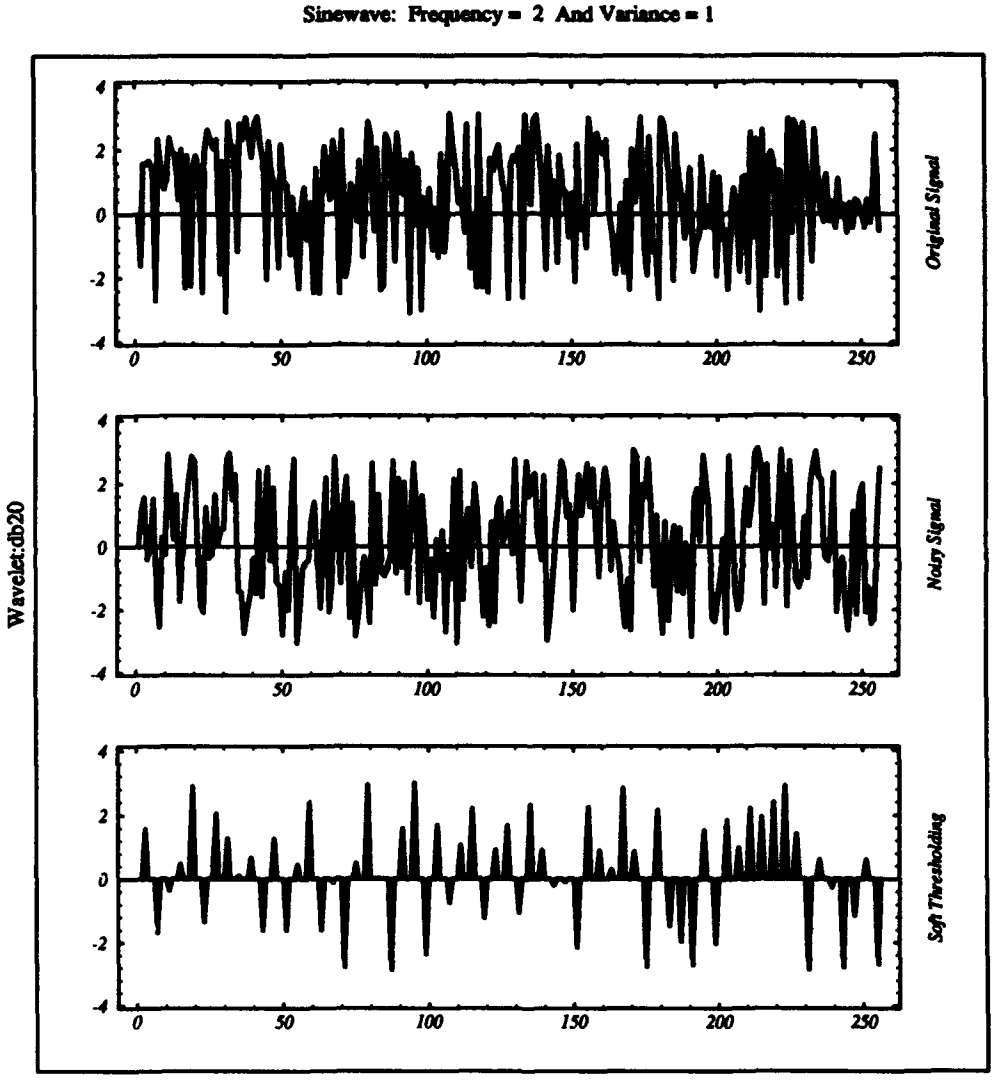

Figure C.6 Phase of the FFT of the clean sinewave (2Hz) after the STT ($\sigma^2 = 1$).

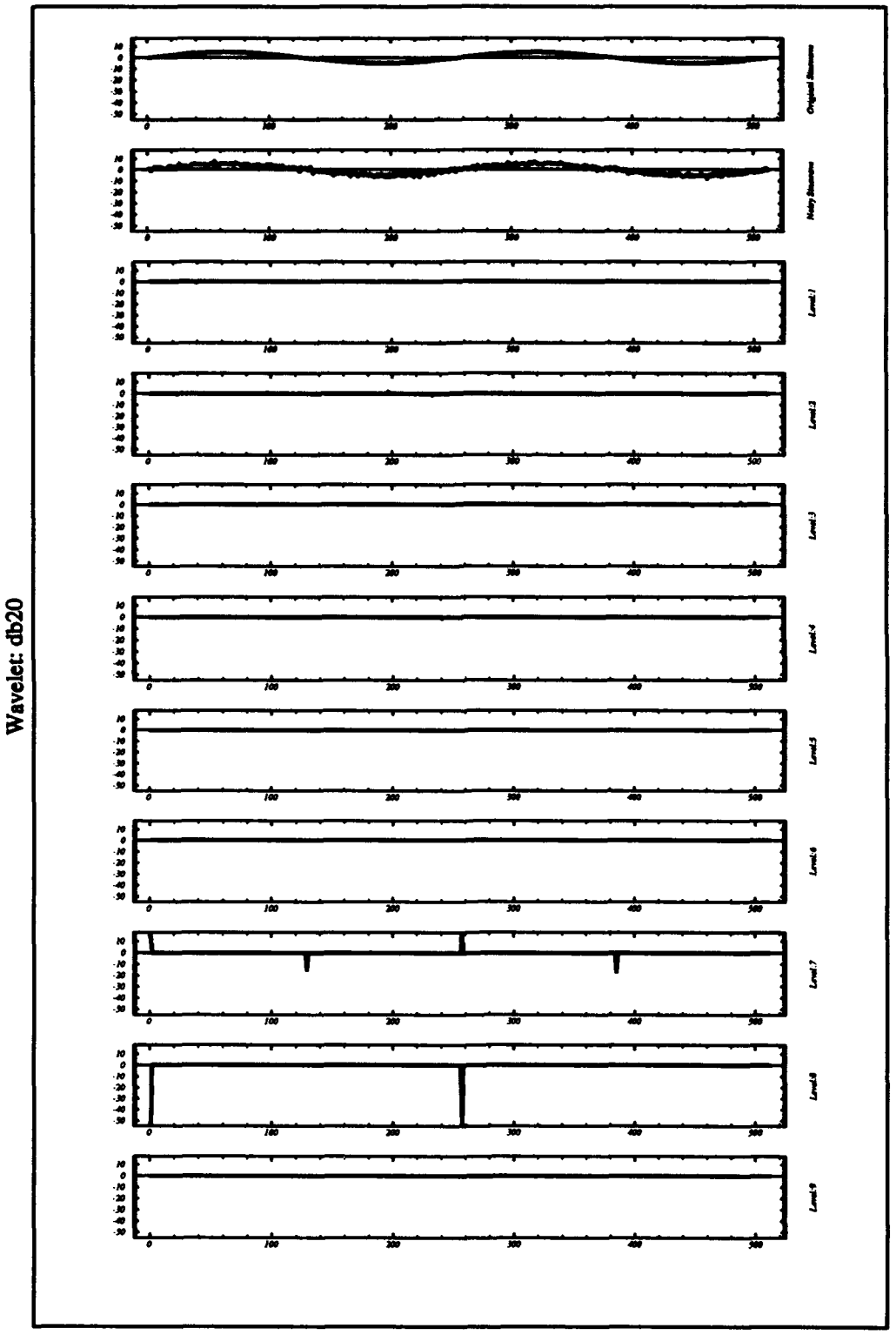

Figure C.7 Details of the processed noisy sinewave (2Hz) after the STT ($\sigma^2 = 1$).

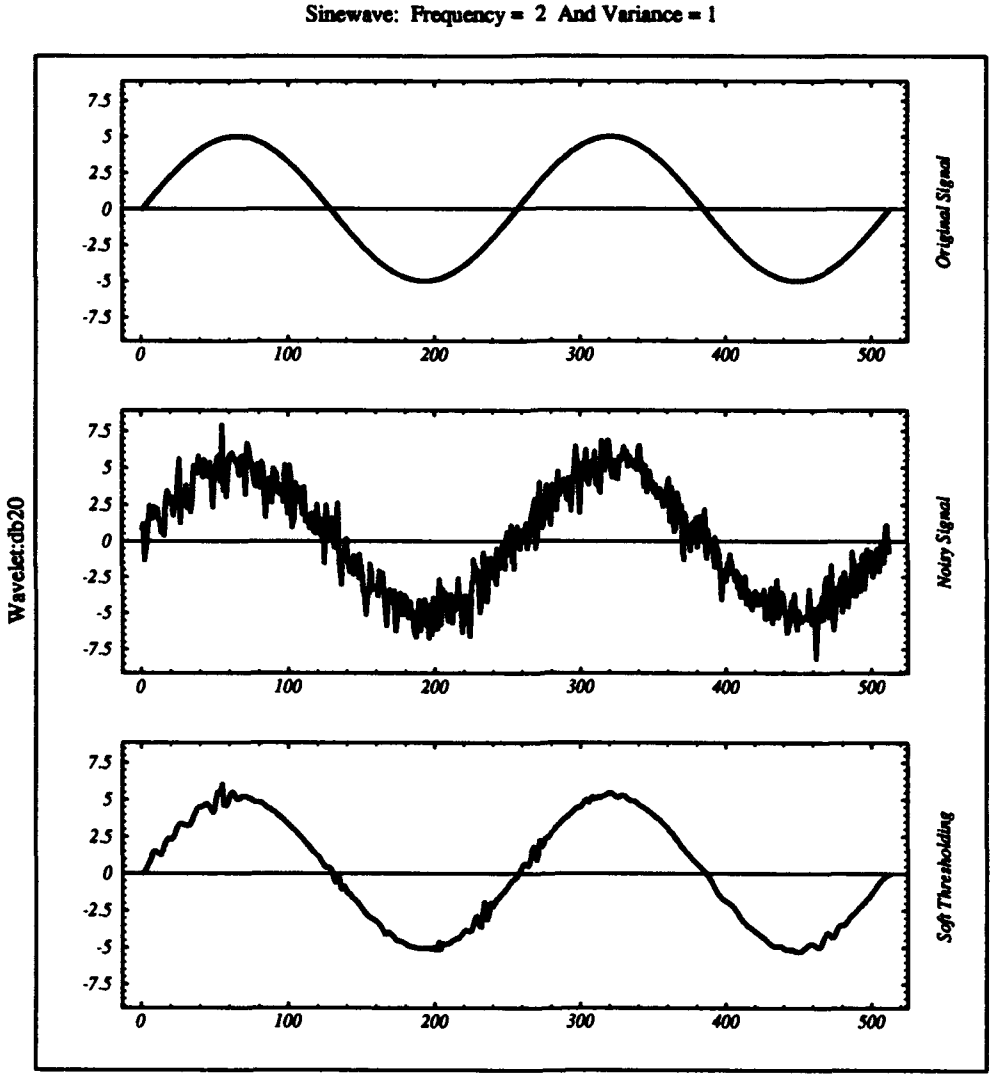

Figure C.8 Noisy sinewave (2Hz) after the STT ($\sigma^2 = 1$).

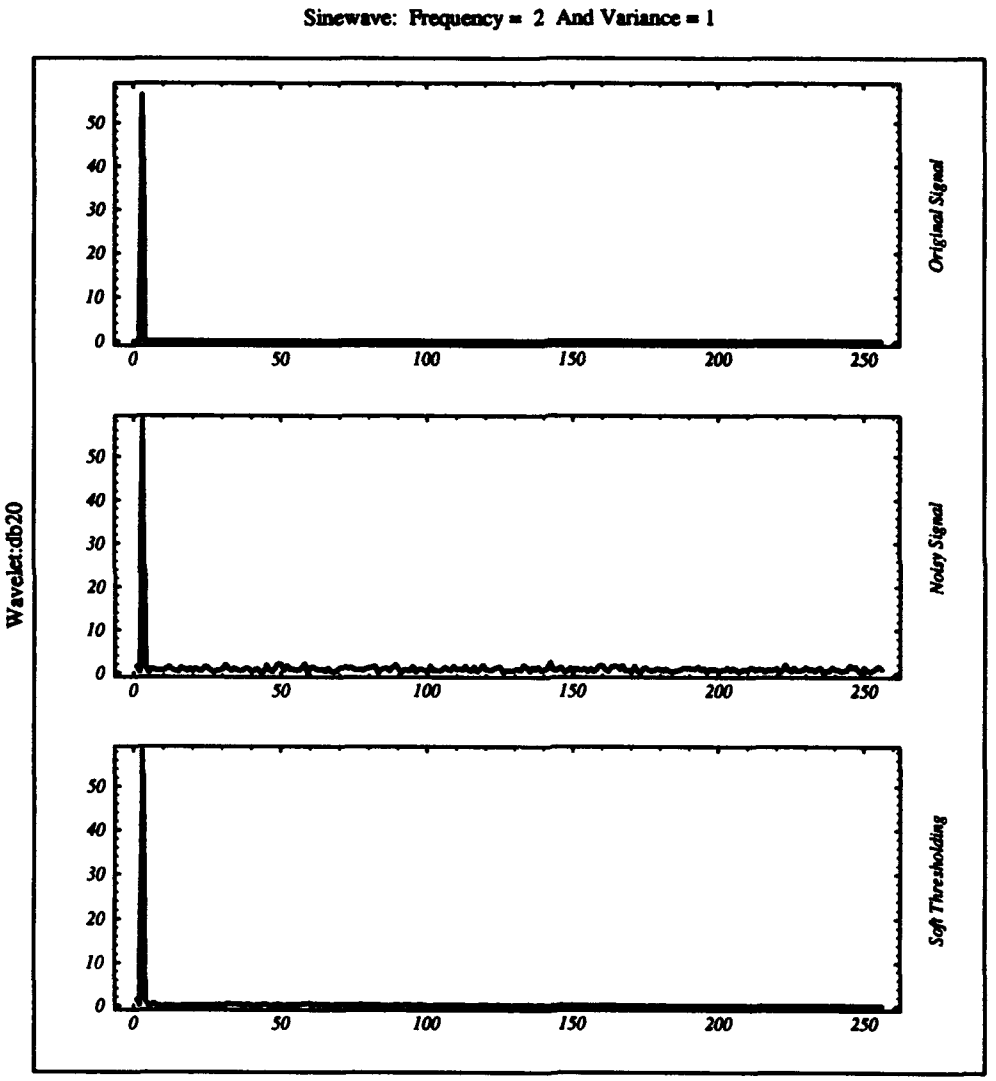

Figure C.9 Amplitude Of the FFT of the noisy sinewave (2Hz) after the STT ($\sigma^2 = 1$).

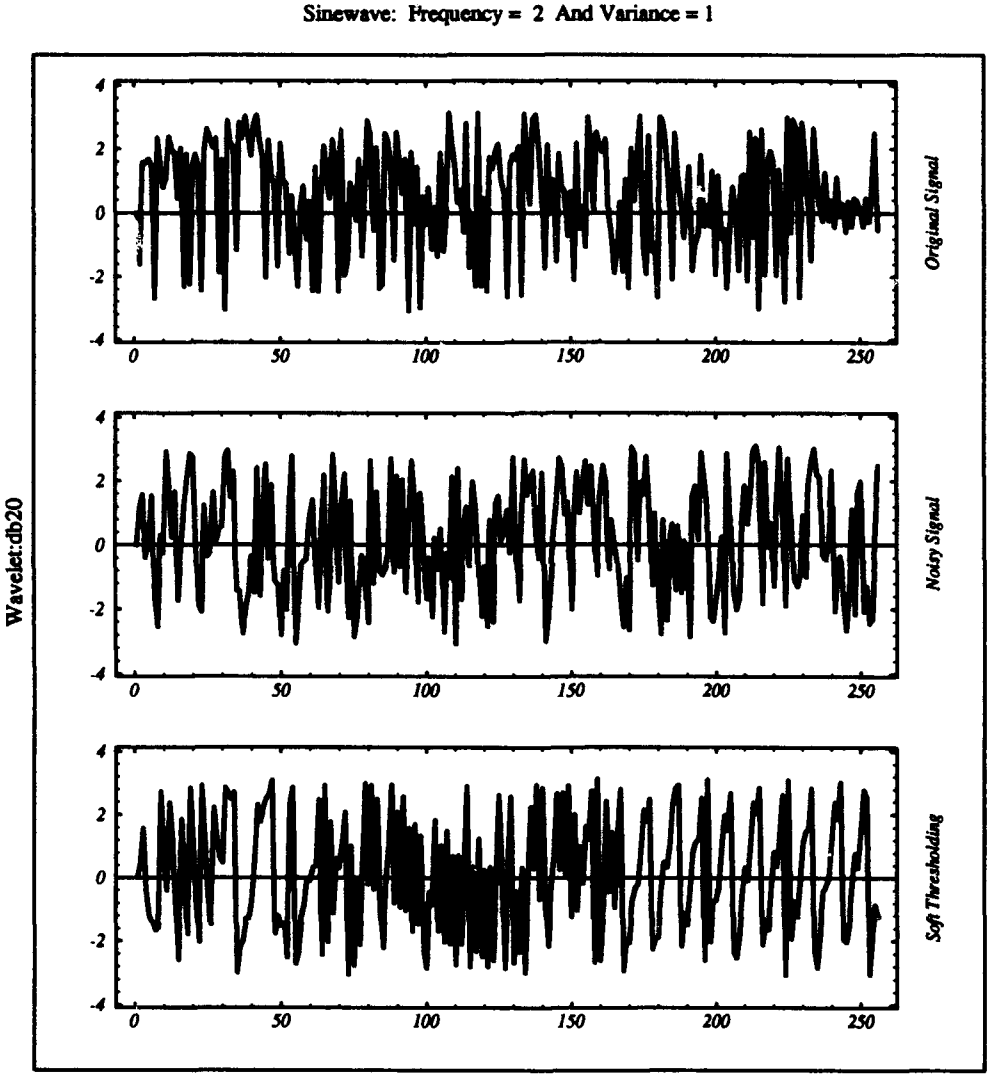

Figure C.10 Phase Of the FFT of the noisy sinewave (2Hz) after the STT ($\sigma^2 = 1$).

Appendix D. Effect Of Wavelet Shrinkage On White Gaussian Noise and Unvoiced Speech

This Appendix contains the plots of a white Gaussian noise signal and an unvoiced speech signal before and after the application of the soft thresholding technique (STT). The de-noising method uses wavelets in the time domain without noisy phase restoration (using db20). 512 samples of a white Gaussian noise signal with zero mean and $\sigma^2 = 1$ was generated. This signal was processed using the STT. Since the SURE function estimates the mean of an independent and normally distributed random signal, the expected result is a signal with 512 zeros (i.e., the noise has zero mean). The white Gaussian plots illustrate the fact that the application of the STT to the white Gaussian noise is very close to zero. Observe, the Fourier transform of the noise has high energy throughout the entire energy spectrum. Also, notice that the high decomposition detail levels (i.g., 1,2, and 3) filter most of the white Gaussian noise.

The second set of plots, deals with both clean and noisy unvoiced speech. The plots illustrate the fact that unvoiced speech is treated as white Gaussian noise. In fact, when using the STT to de-noise a clean unvoiced speech signal (similar to the case of a clean sinewave), the result is a signal with zeros everywhere (see figure D.11). Notice that the effects of the STT on a noisy unvoiced speech signal are similar to the effects of the STT on white Gaussian noise. We conclude then, that the noisy unvoiced speech data has characteristics similar to those of white Gaussian noise. In order to prevent loosing all the unvoiced as well as the silent speech portions, we chose not to process these portions. One important observation is that both the soft and hard thresholding techniques (STT and HTT) can be used as detectors for voiced, unvoiced, and silent speech segments.

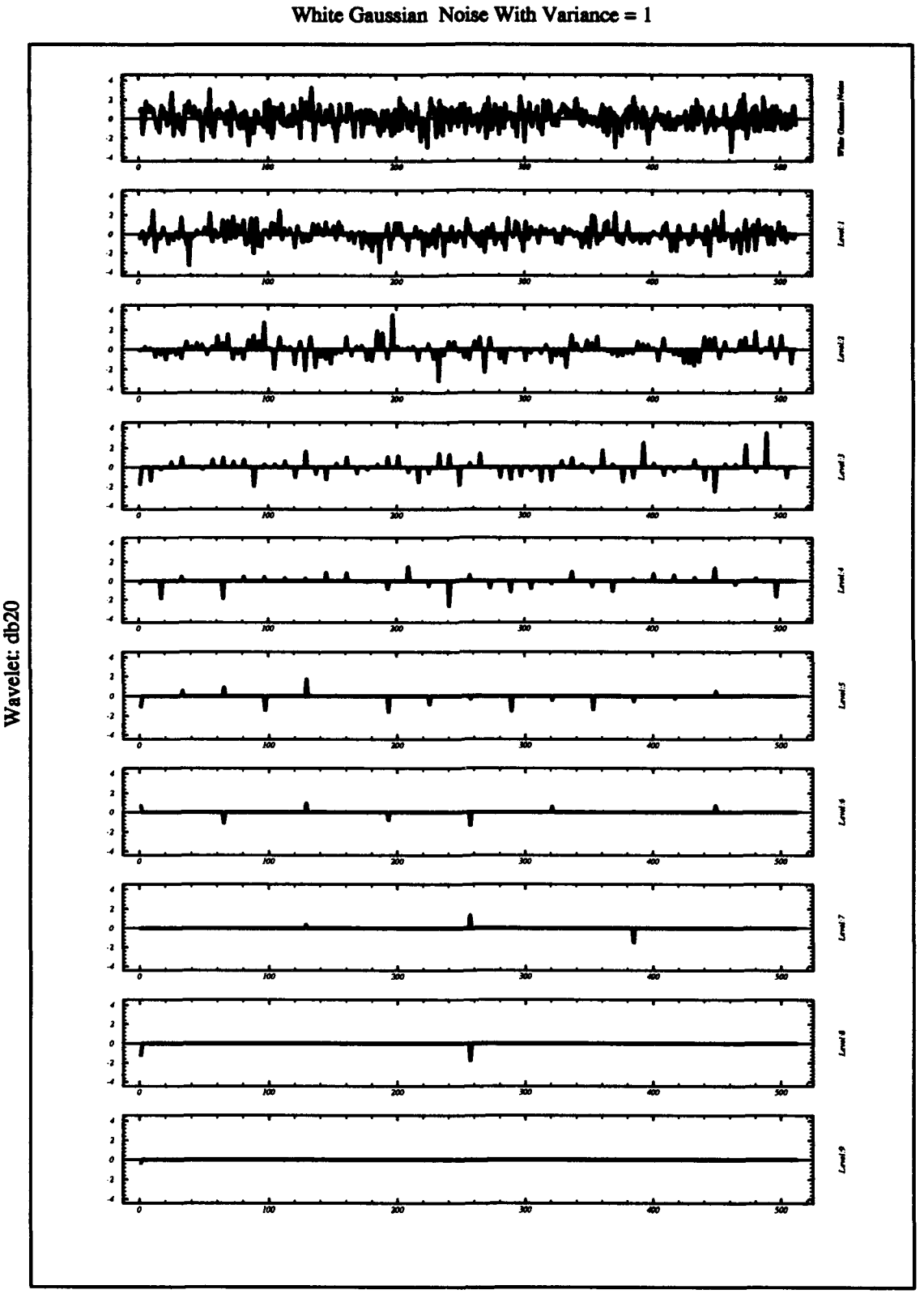

Figure D.1 Details of the white Gaussian noise ($\sigma^2 = 1$).

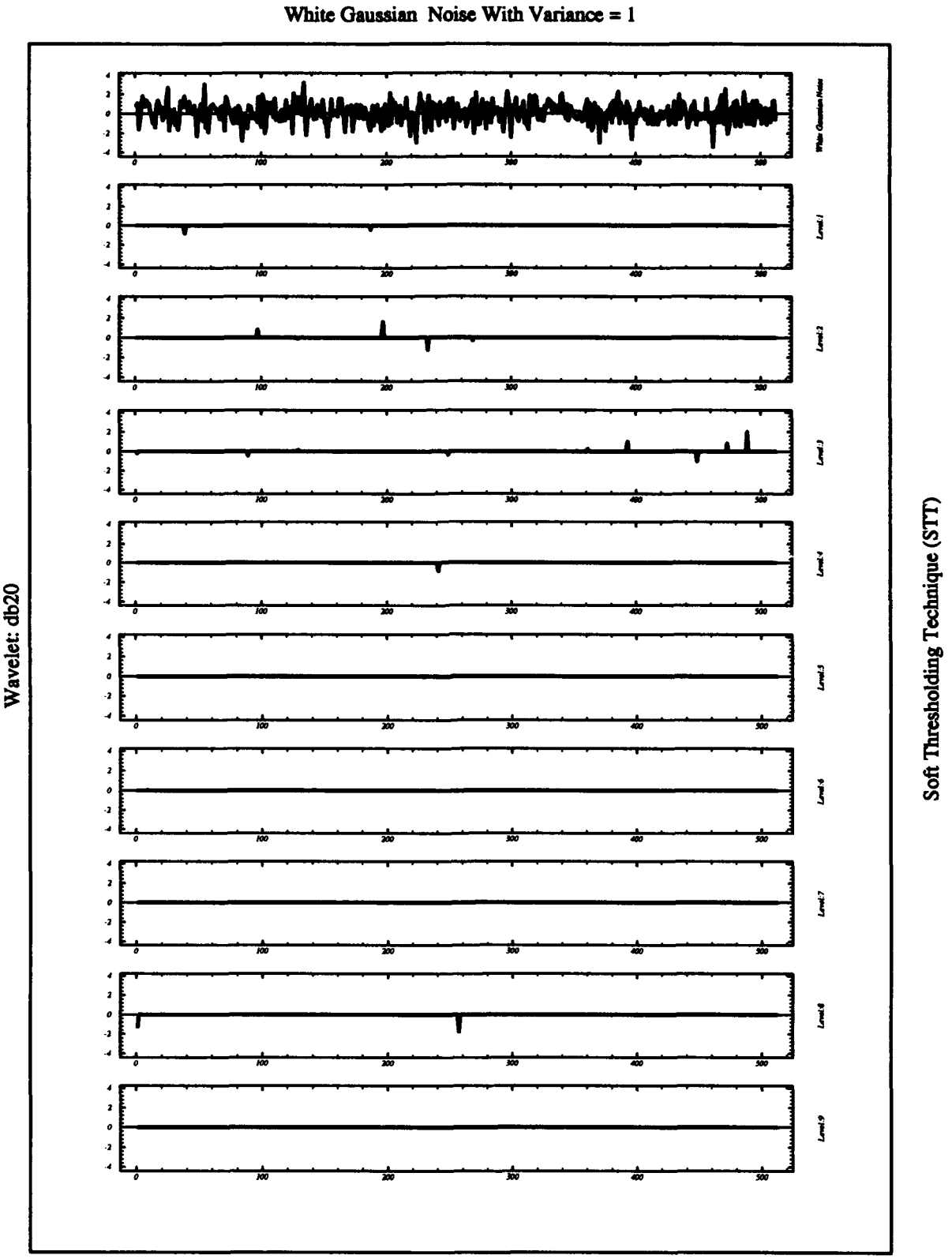

Figure D.2 Details of the processed white Gaussian noise after the STT ($\sigma^2 = 1$).

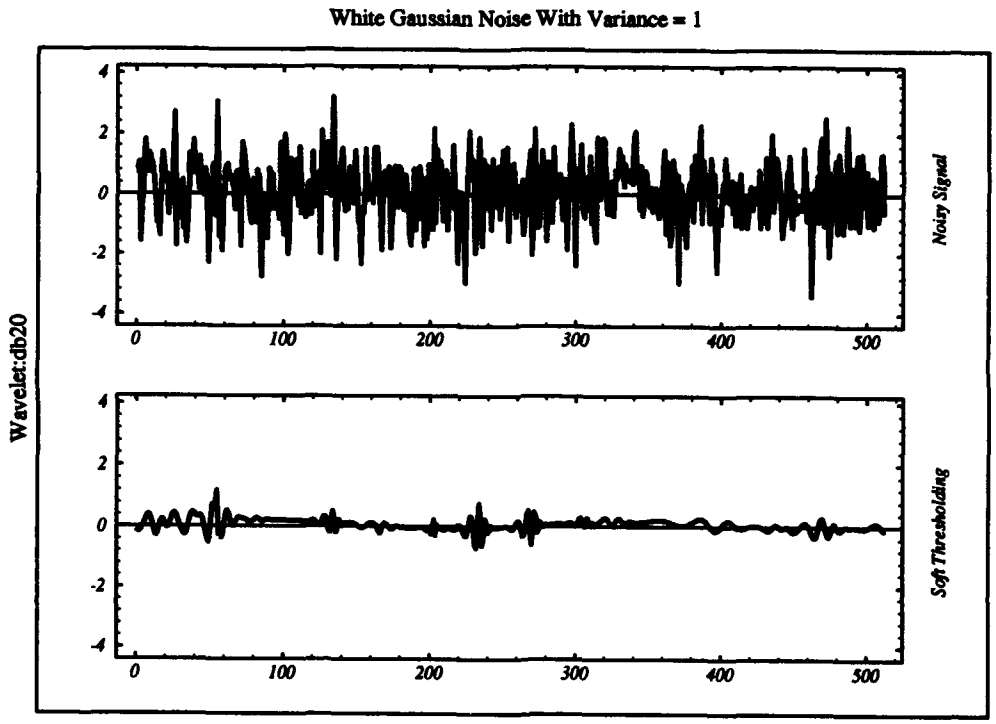

Figure D.3 White Gaussian noise after the STT ($\sigma^2 = 1$).

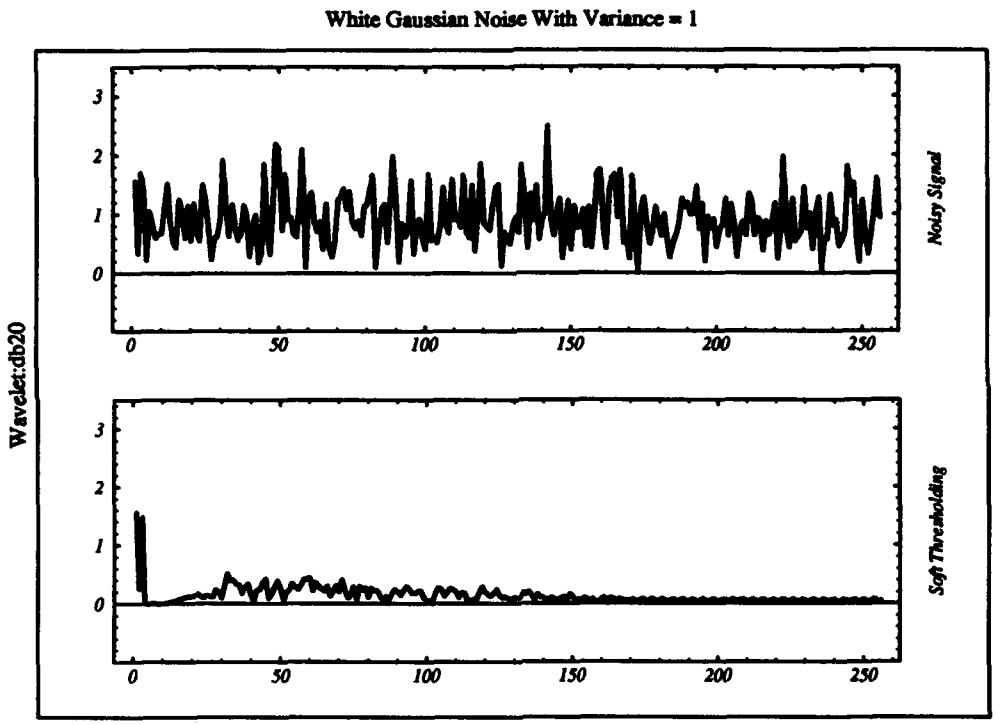

Figure D.4 Amplitude Of the FFT of the white Gaussian noise after the STT ($\sigma^2 = 1$).

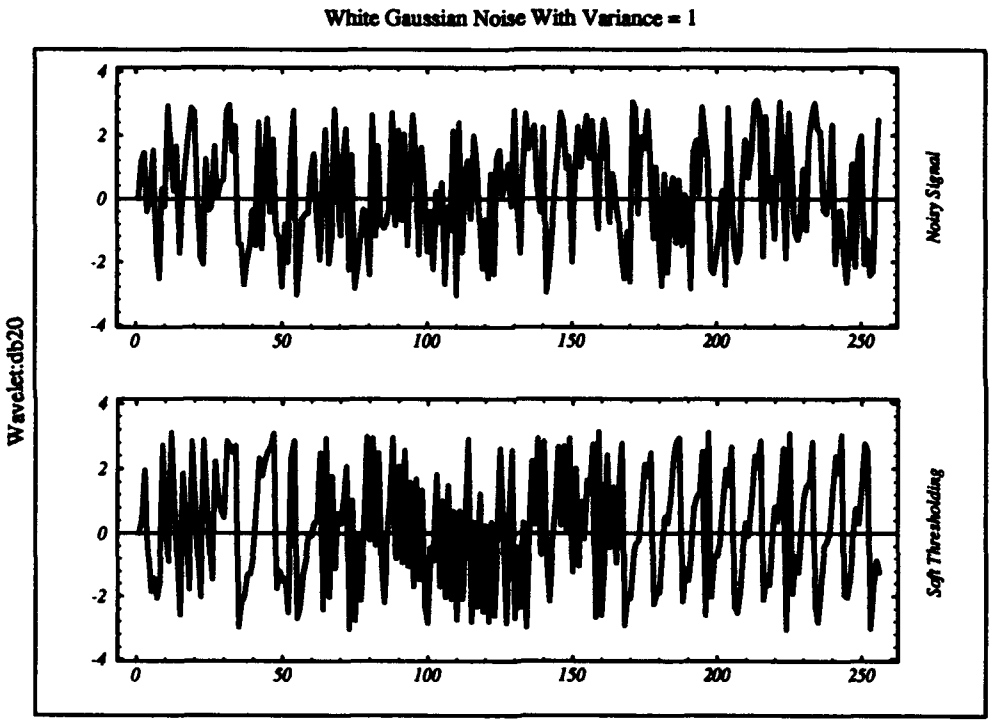

Figure D.5 Phase Of the FFT of the white Gaussian noise after the STT ($\sigma^2 = 1$).

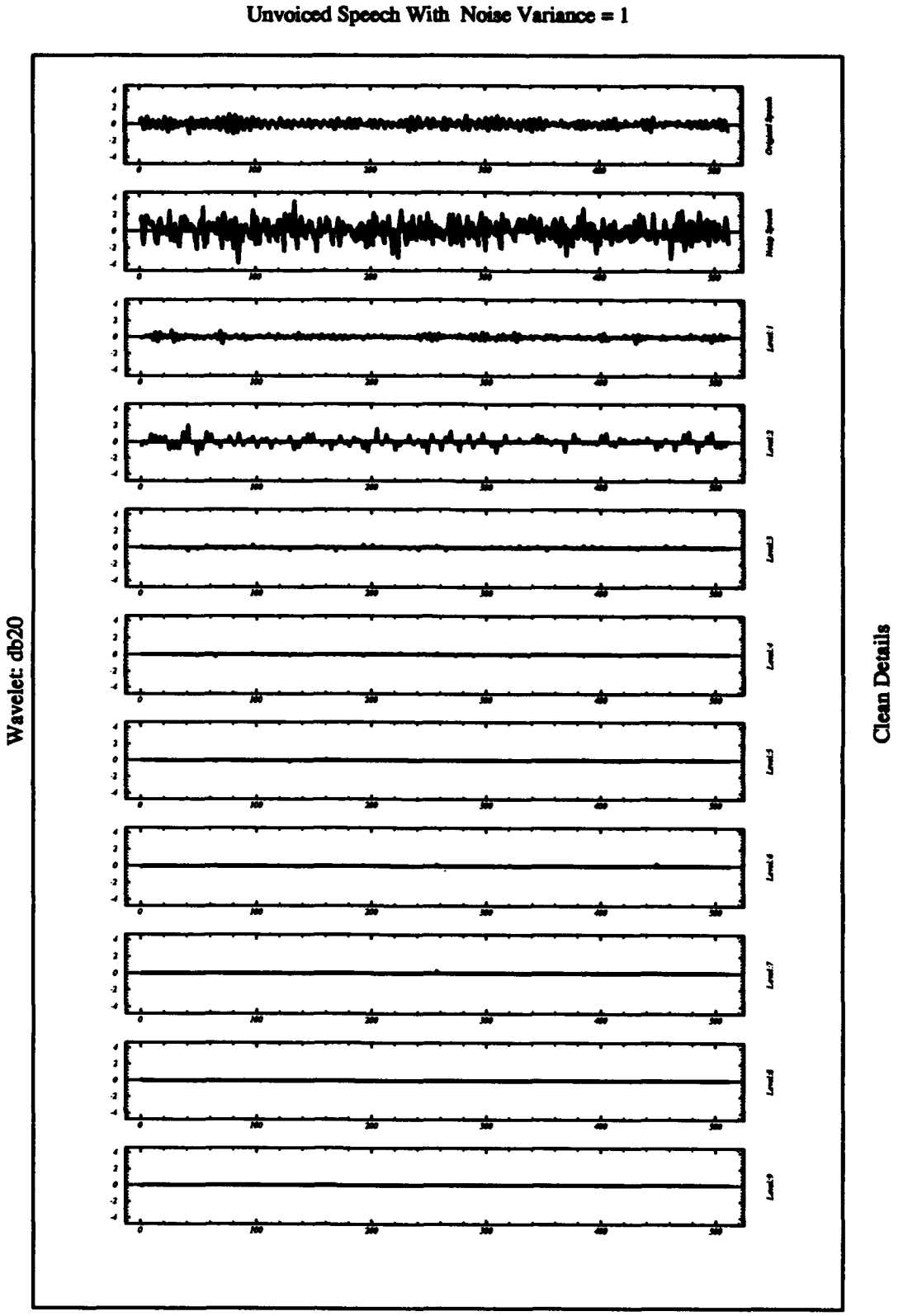

Figure D.6 Details of the clean unvoiced speech.

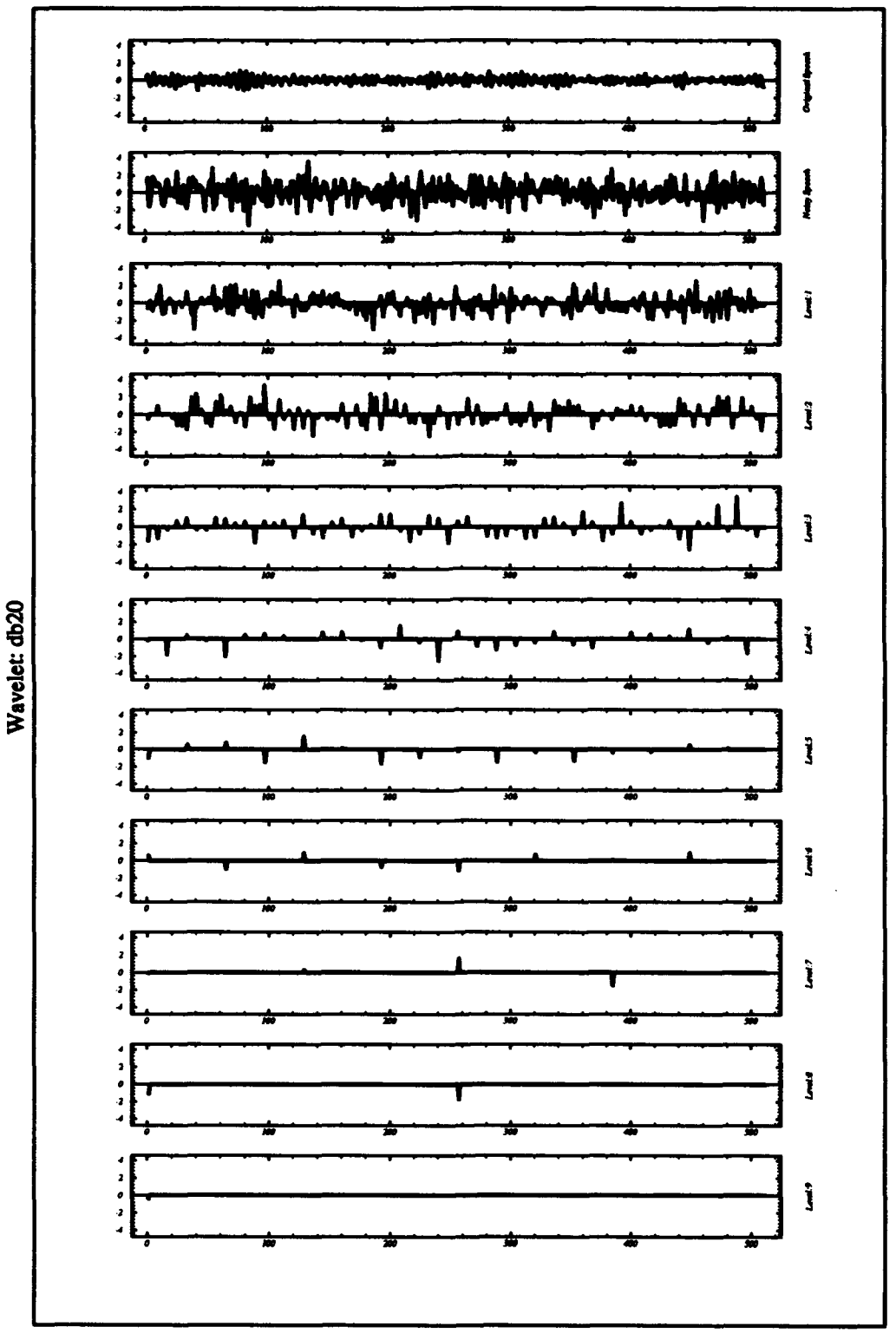

Figure D.7 Details of the noisy unvoiced speech.

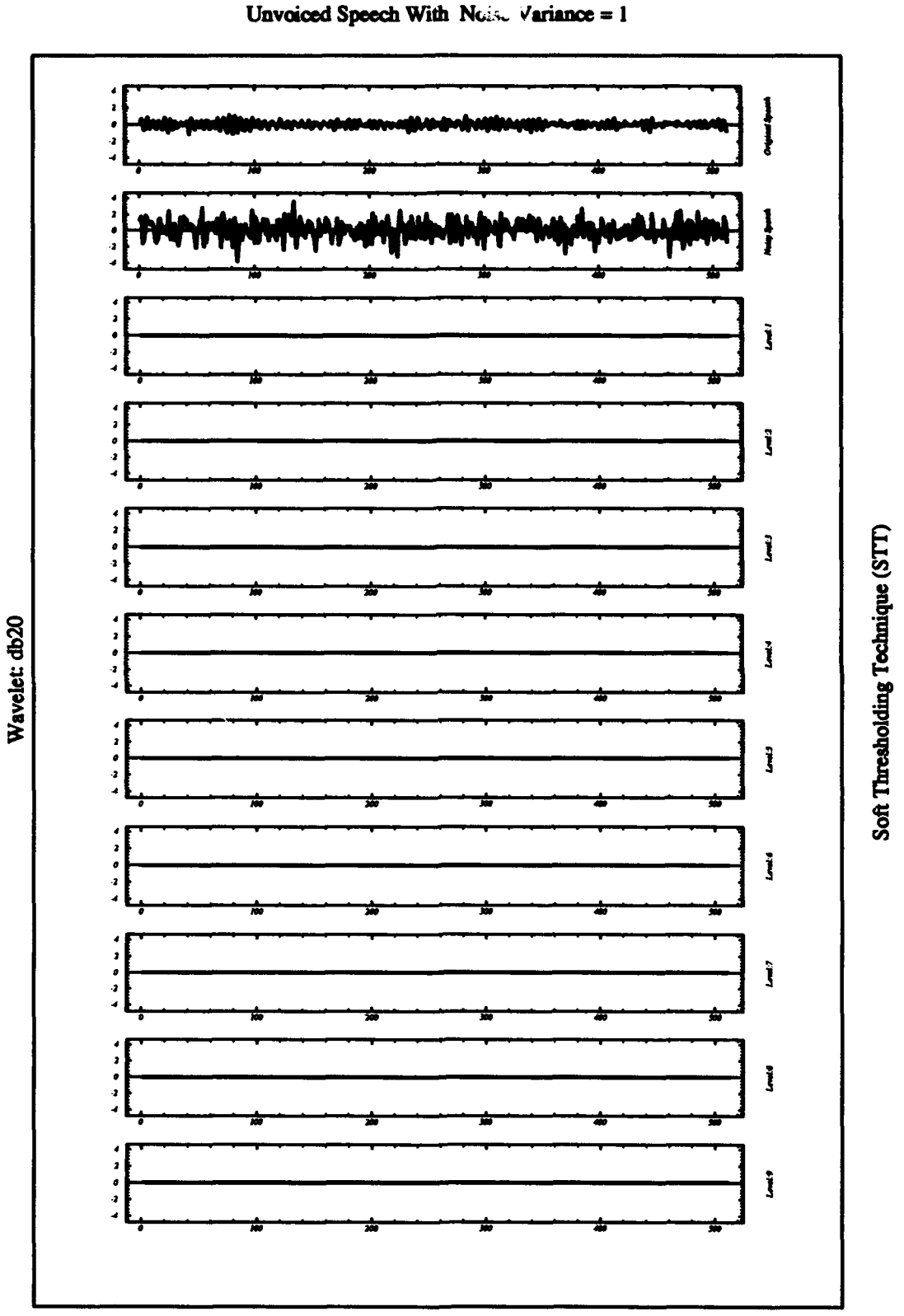

Figure D.8 Details of the processed clean unvoiced speech after the STT ($\sigma^2 = 1$).

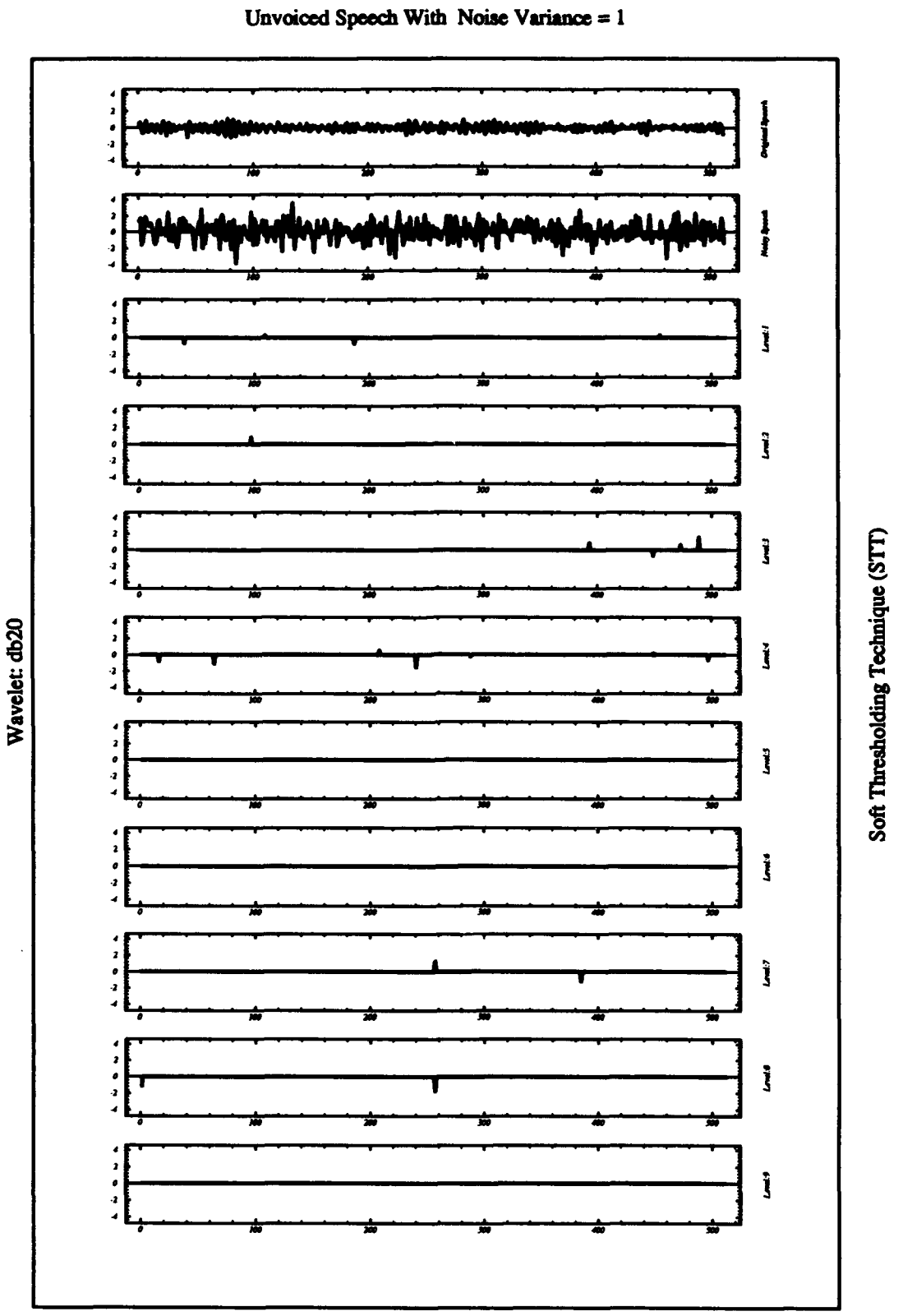

Figure D.9 Details of the processed noisy unvoiced speech after the STT ($\sigma^2 = 1$).

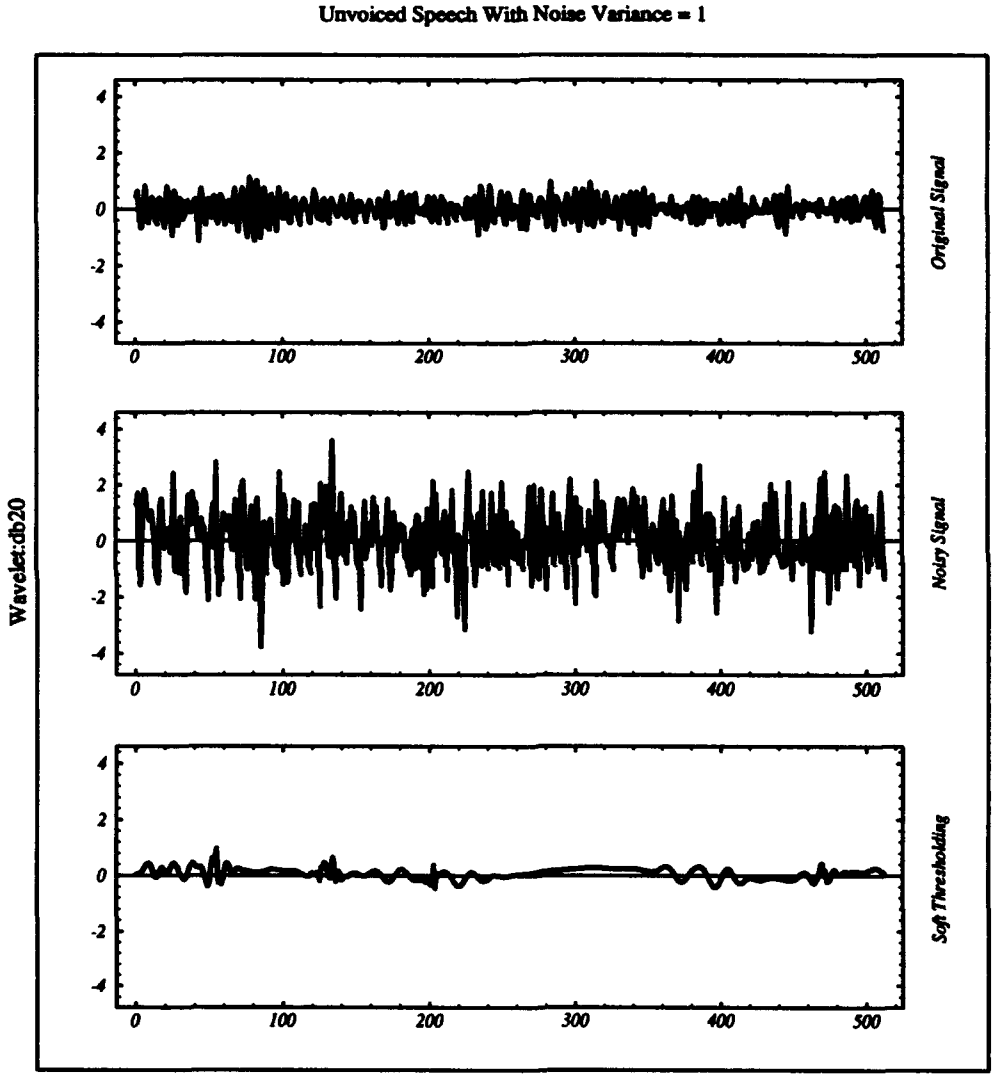

Figure D.10 Noisy unvoiced speech after the STT ($\sigma^2 = 1$).

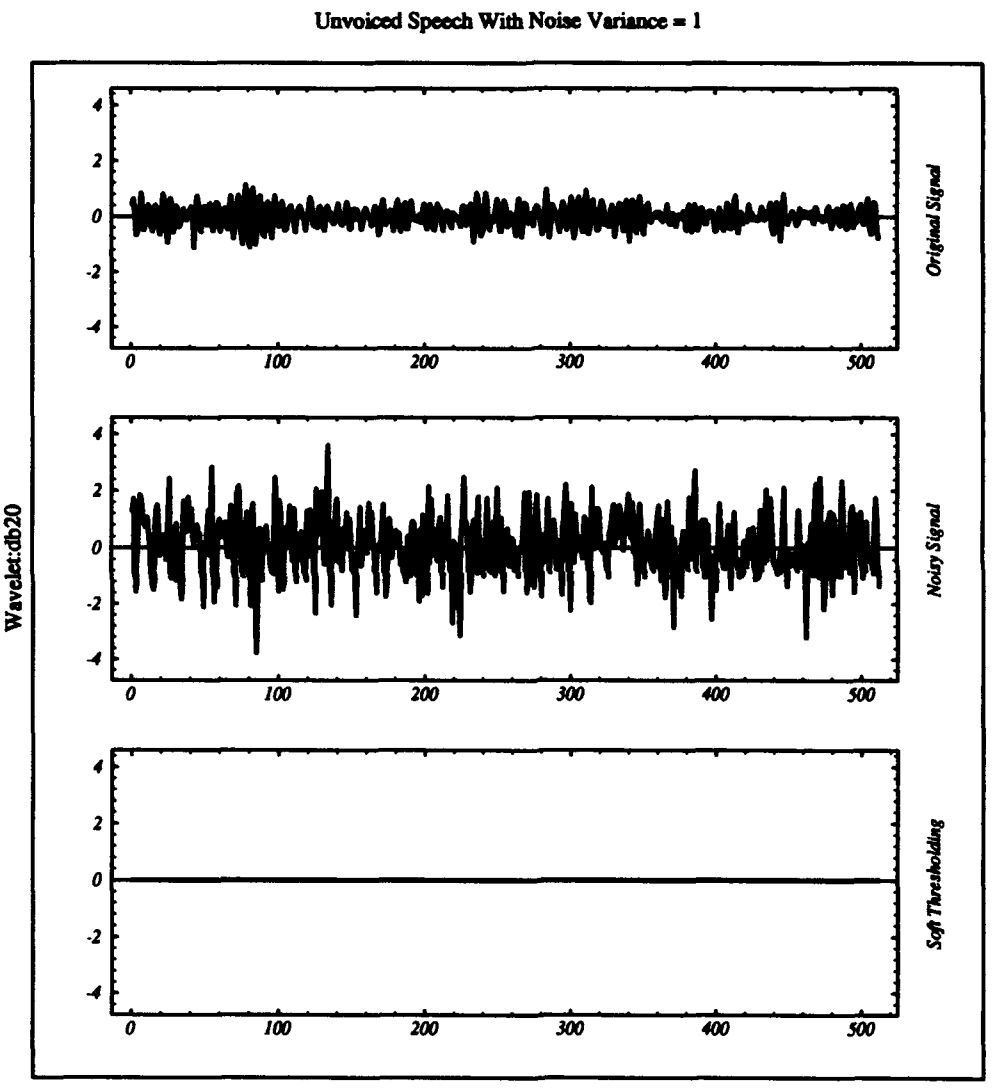

Figure D.11 Clean unvoiced speech after the STT ($\sigma^2 = 1$).

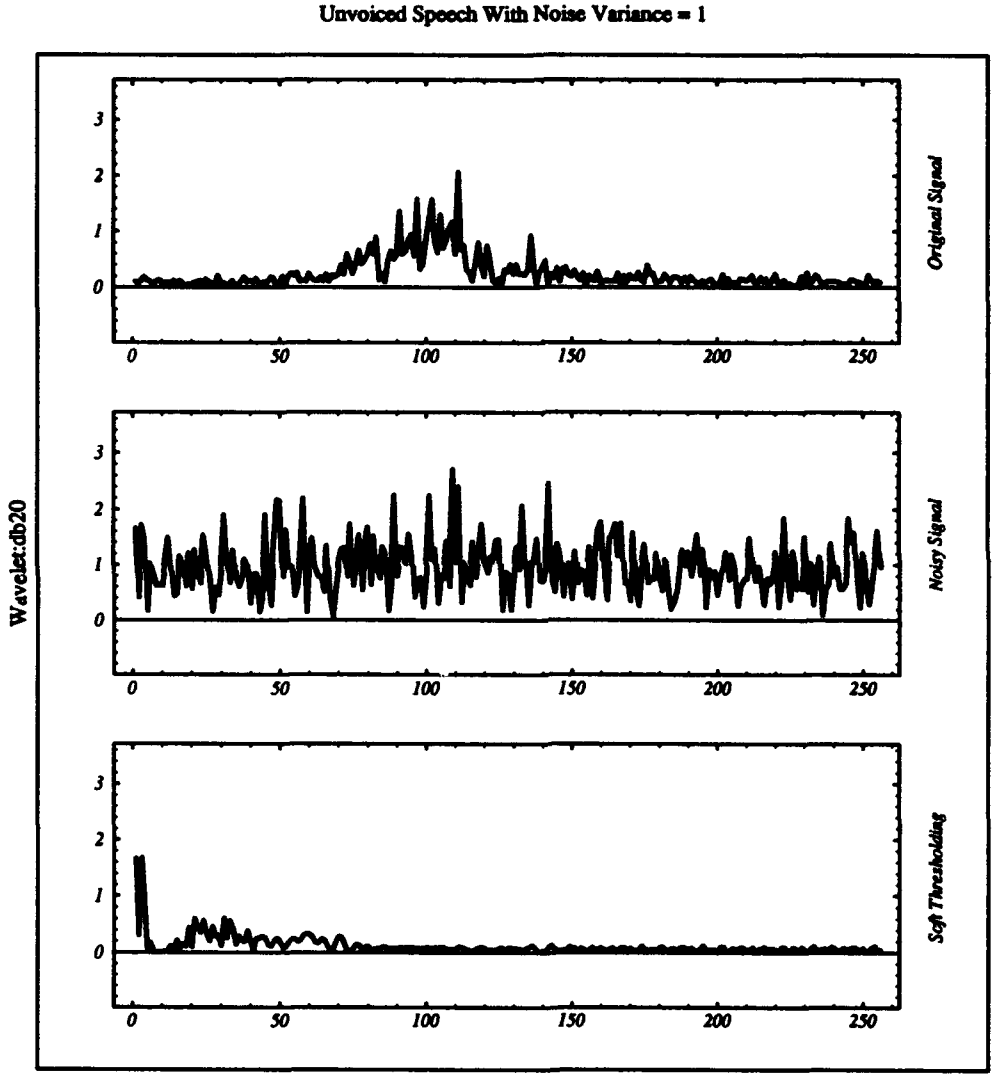

Figure D.12 Amplitude Of the FFT of the noisy unvoiced speech after the STT ($\sigma^2 = 1$).

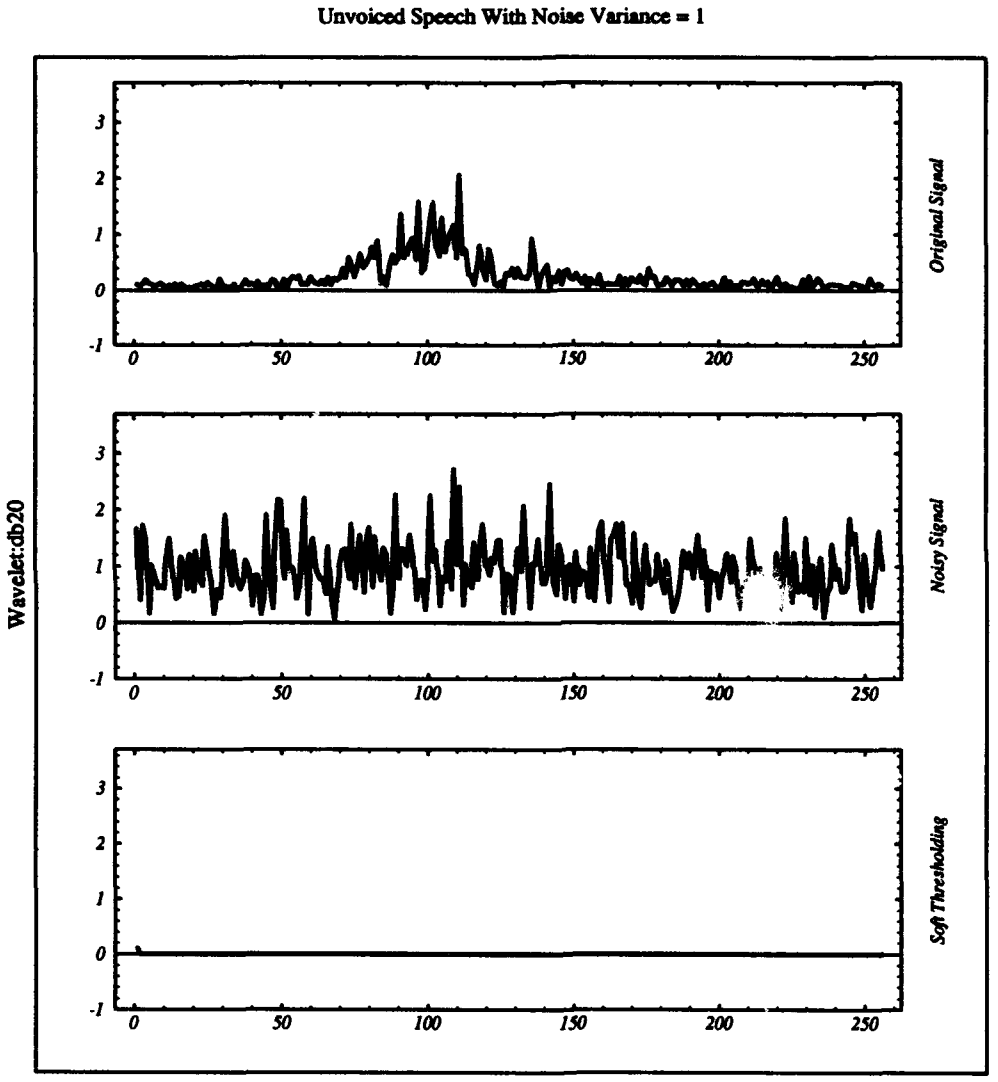

Figure D.13 Amplitude of the FFT of the clean unvoiced speech after the STT ($\sigma^2 = 1$).

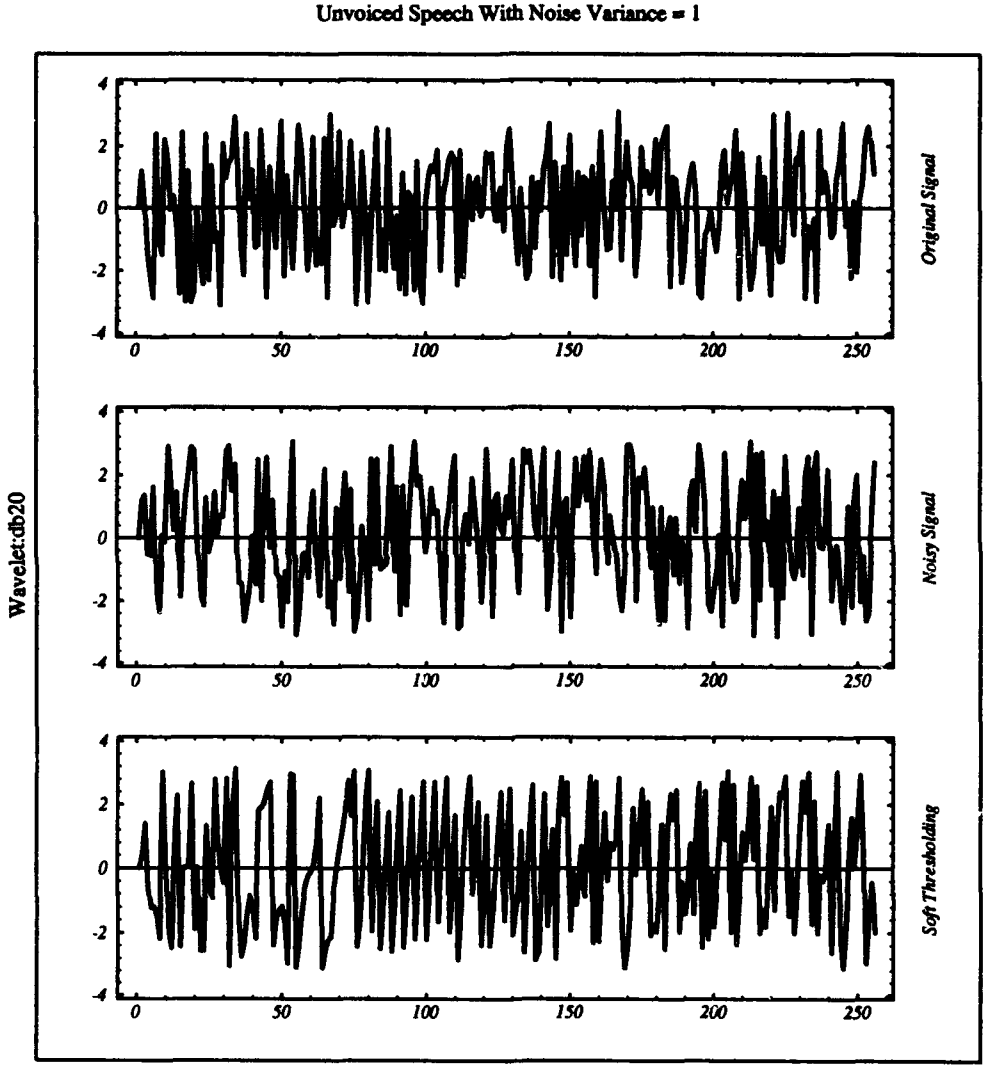

Figure D.14 Phase Of the FFT of the noisy unvoiced speech after the STT ($\sigma^2 = 1$).

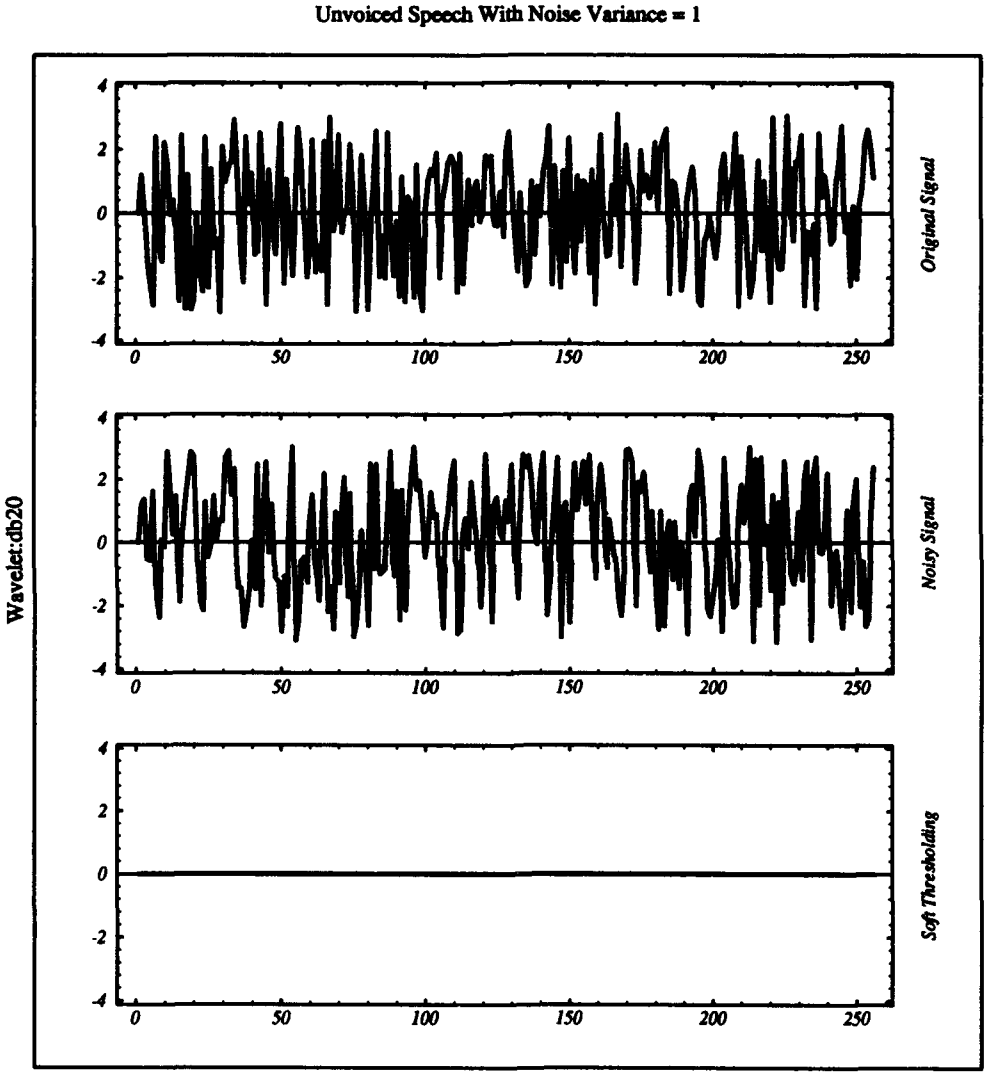

Figure D.15 Phase of the FFT of the clean unvoiced speech after the STT ($\sigma^2 = 1$).

Appendix E. Effect Of Wavelet Shrinkage On Voiced Speech

This Appendix illustrates the effects of applying the STT to both a noisy and a clean voiced speech segments. The wavelet decomposition of the clean voiced speech segment shows high energy details at the coarser levels of decomposition (i.e., levels 4, 5, and 6), while the finer levels of decomposition (i.e., levels 1 and 2) have little or no high energy details at all (see figure E.1). On the other hand, the noisy version (clean voiced speech and noise with variance $\sigma^2 = 1$) of this voiced speech signal, shows high energy at all detail levels; especially levels 1 and 2 (see figure E.2). The effects of the STT on both the clean and noisy voiced speech signals is that most of the high frequency details are eliminated (see figures E.3 and E.4). Observe that the reconstruction of both signals, the STT processed clean voiced speech signal and the STT processed noisy voiced speech signal, are very close to the original voiced speech signal. The amplitude of the Fourier transforms of the reconstructed signals show very little high frequency components. Finally, notice the effects of the non-linear processing on the phase (see figures E.10 and E.9).

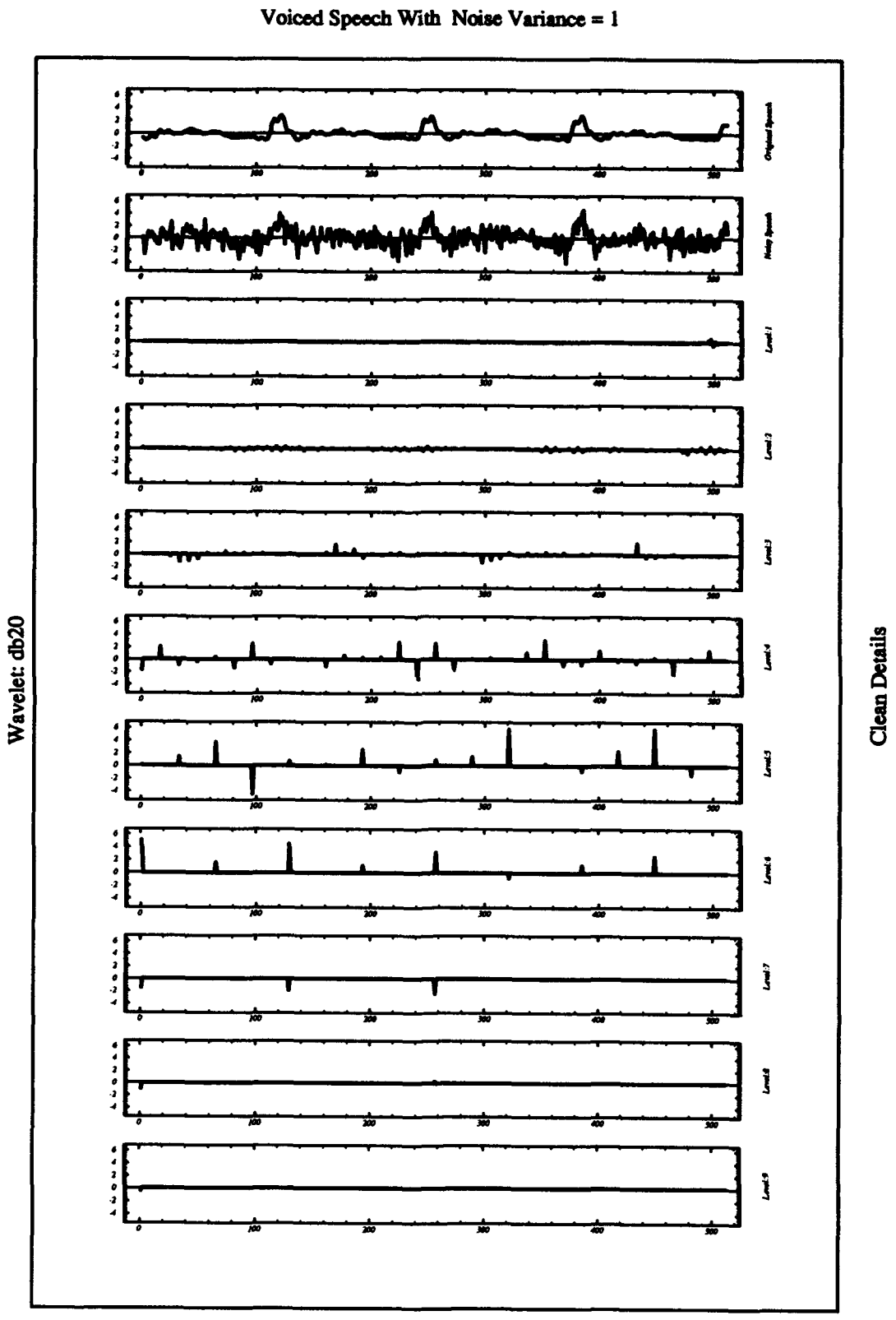

Figure E.1 Details of the clean voiced speech.

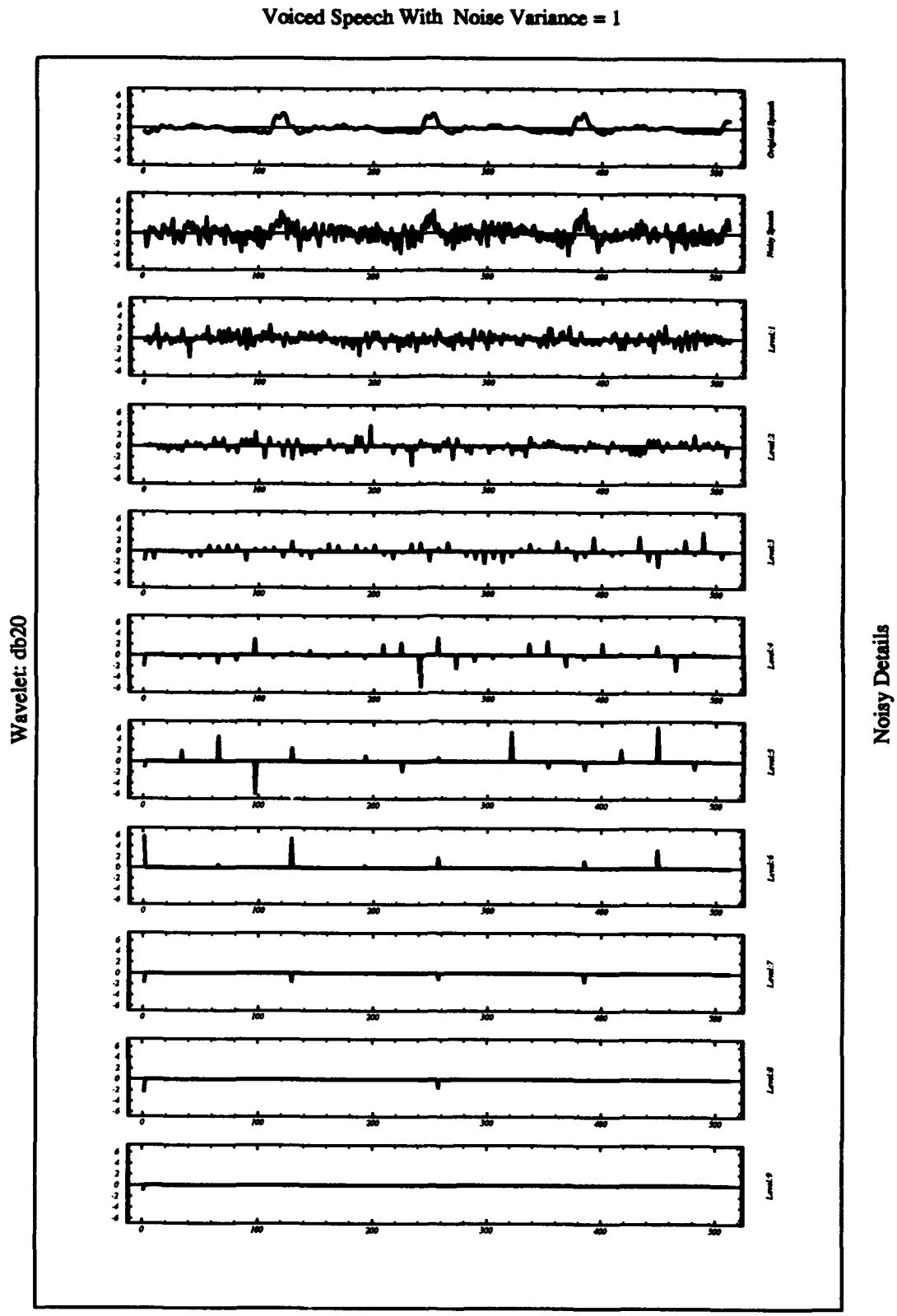

Figure E.2 Details of the noisy voiced speech.

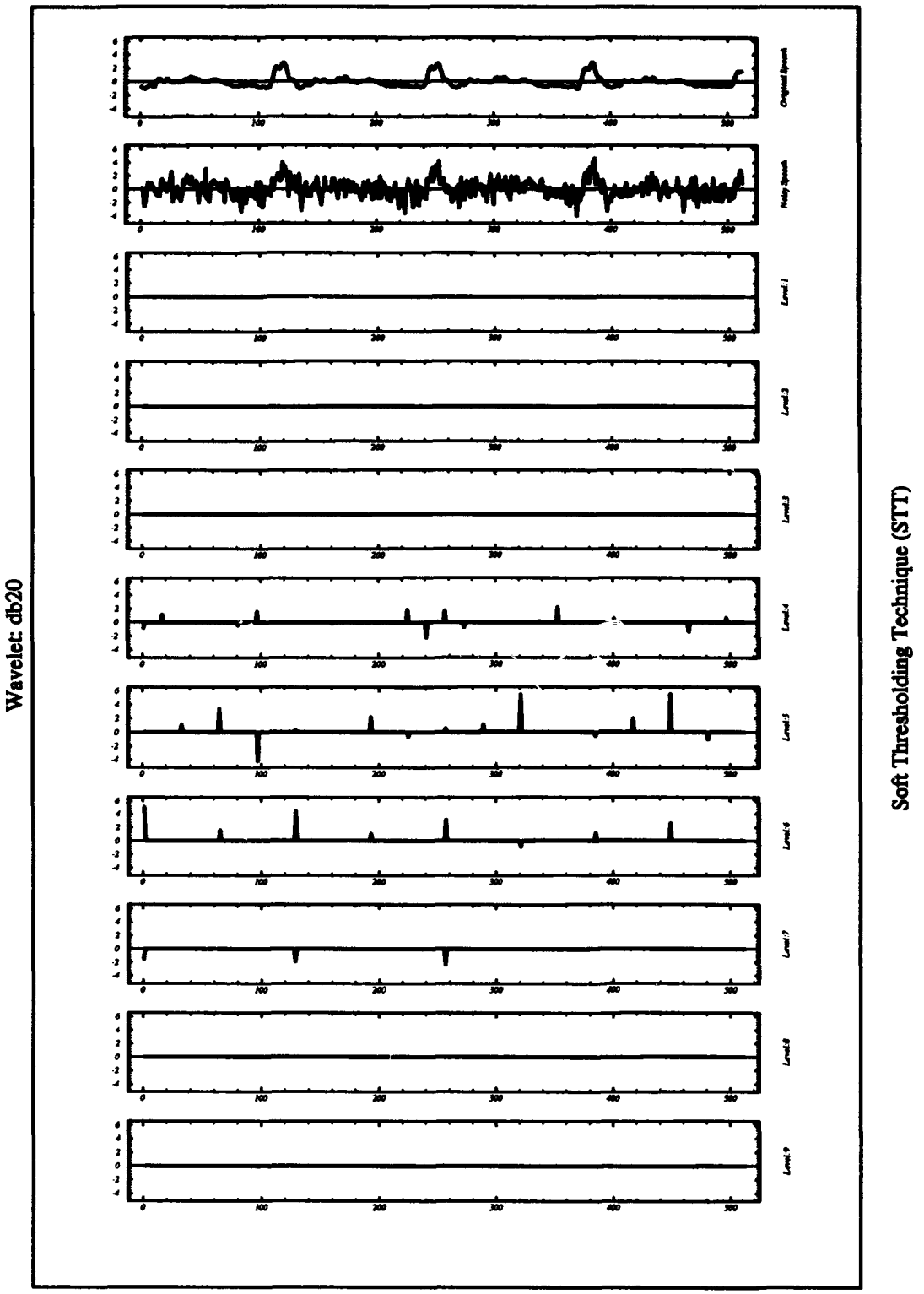

Figure E.3 Details of the processed clean voiced speech after the STT ($\sigma^2 = 1$).

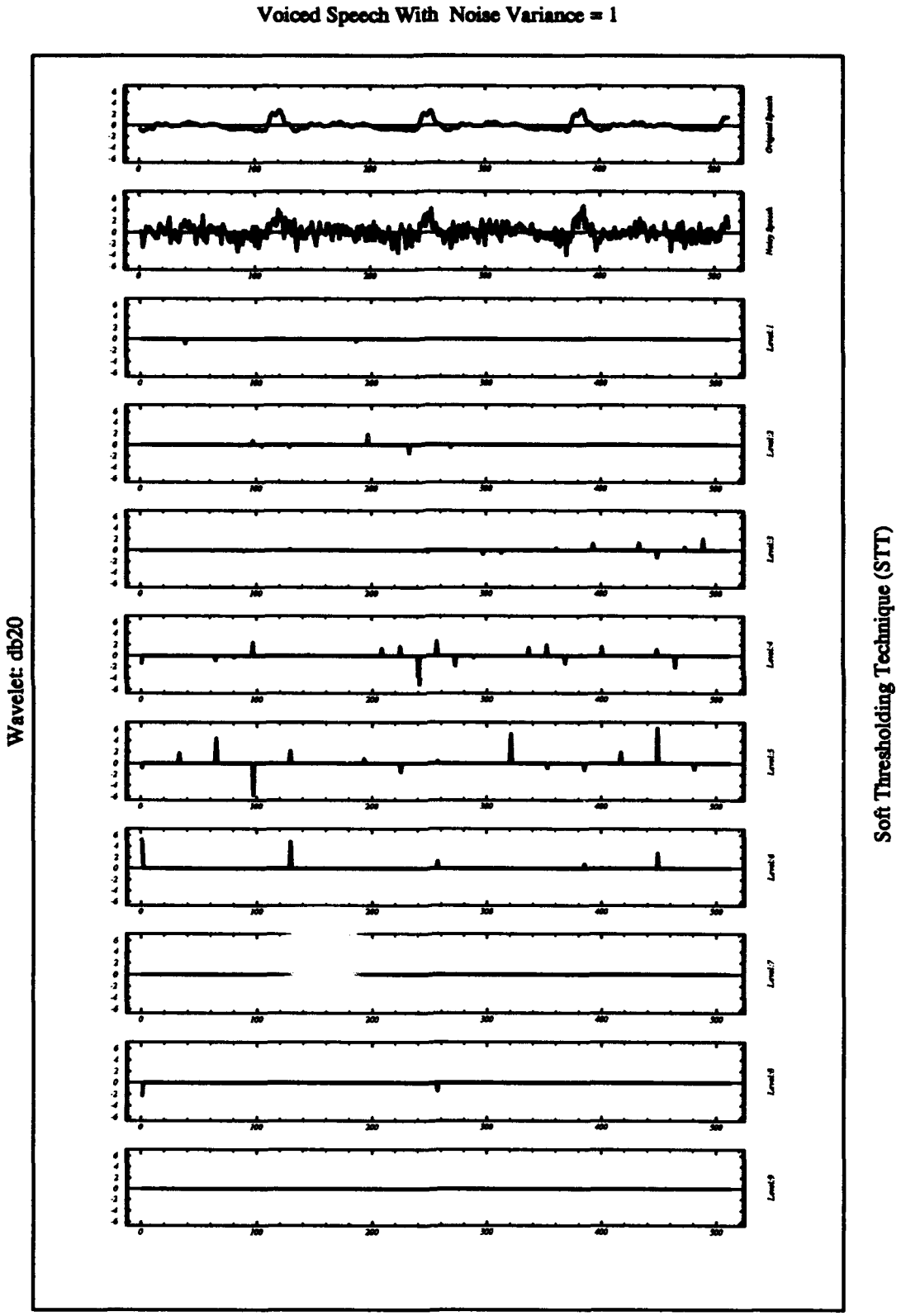

Figure E.4 Details of the processed noisy voiced speech after the STT ($\sigma^2 = 1$).

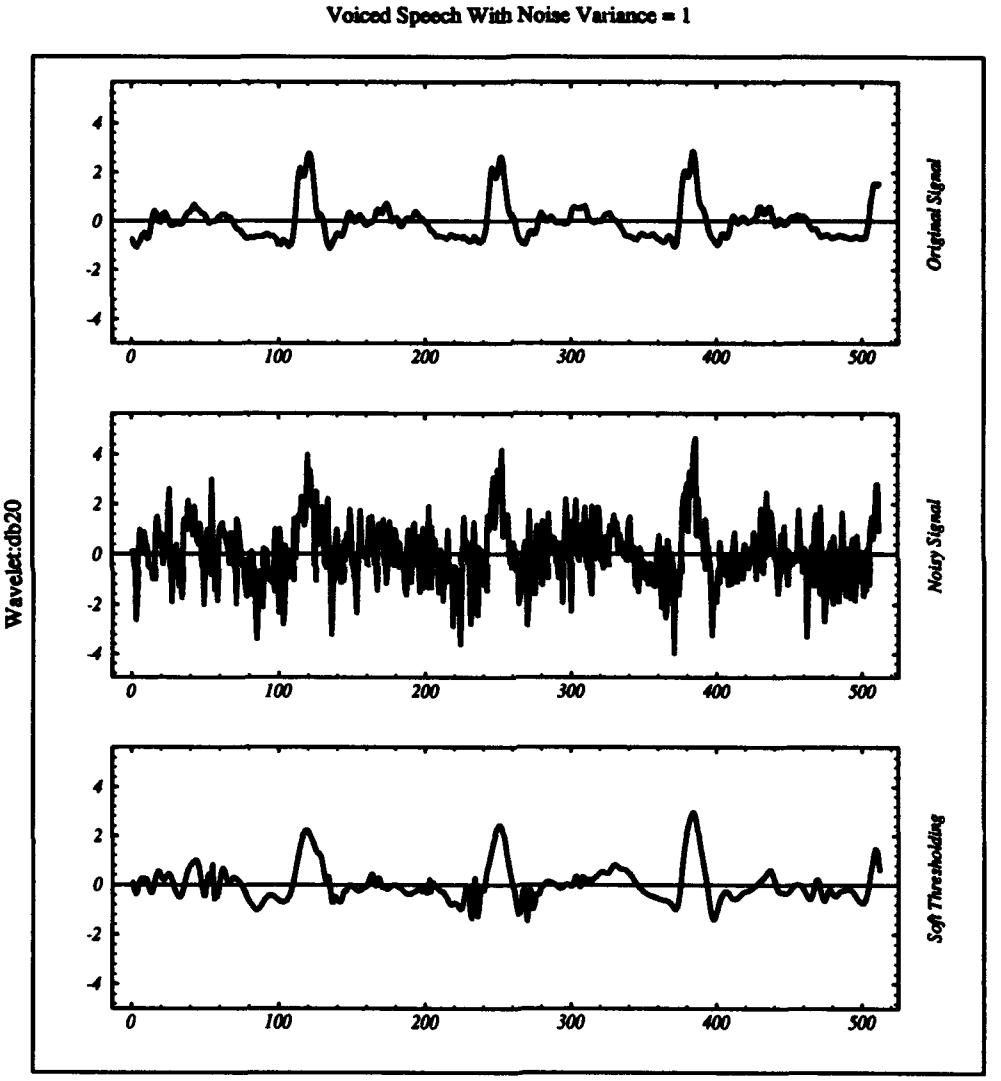

Figure E.5 Noisy voiced speech after the STT ($\sigma^2 = 1$).

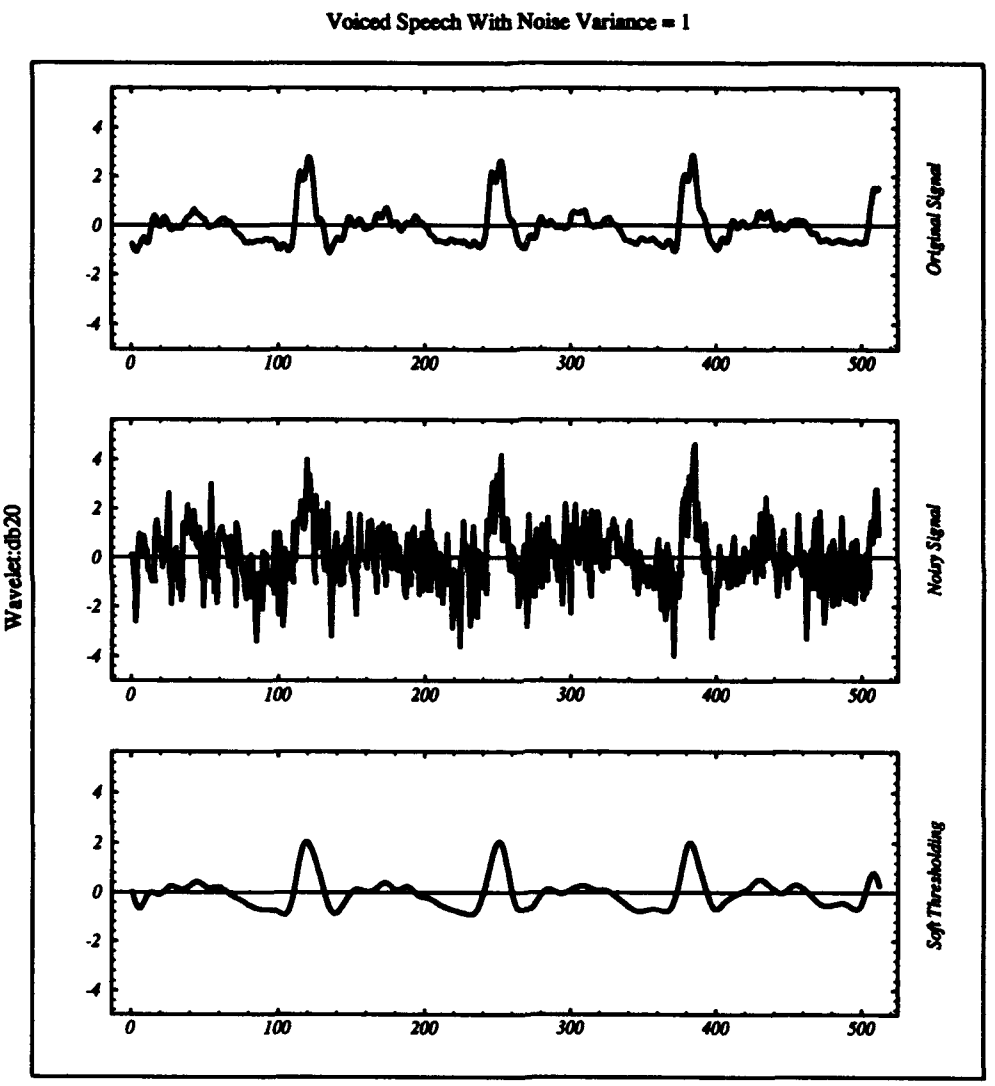

Figure E.6 Clean voiced speech after the STT ($\sigma^2 = 1$).

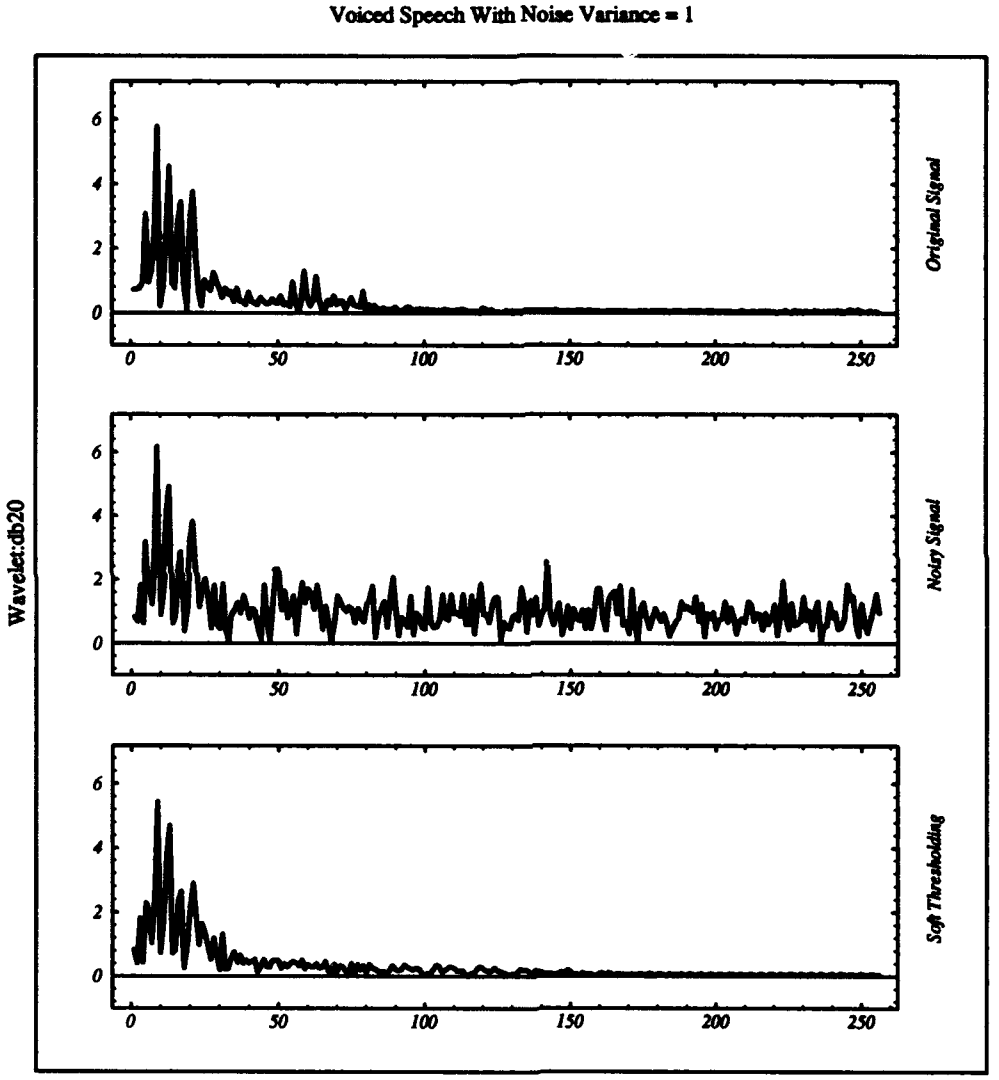

Figure E.7 Amplitude Of the FFT of the noisy voiced speech after the STT ($\sigma^2 = 1$).

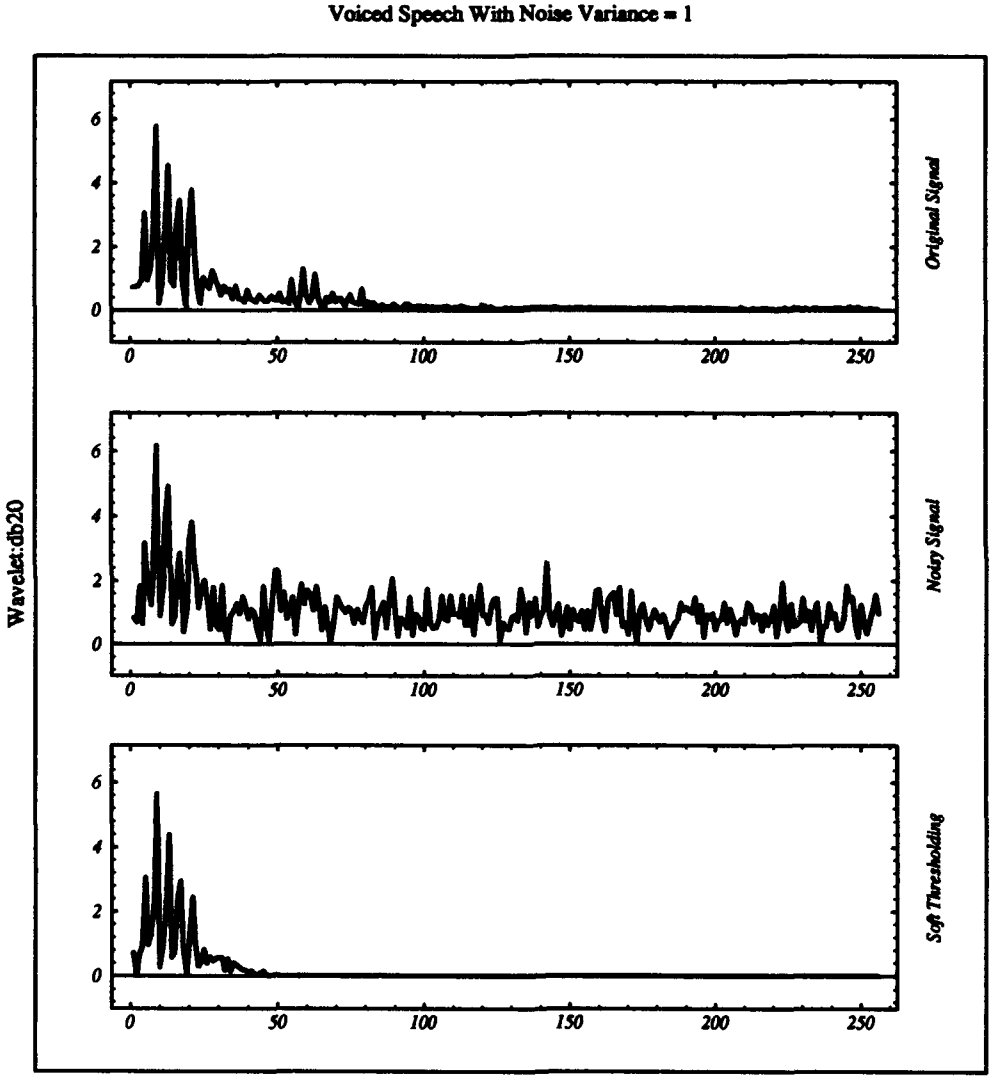

Figure E.8 Amplitude of the FFT of the clean voiced speech after the STT ($\sigma^2 = 1$).

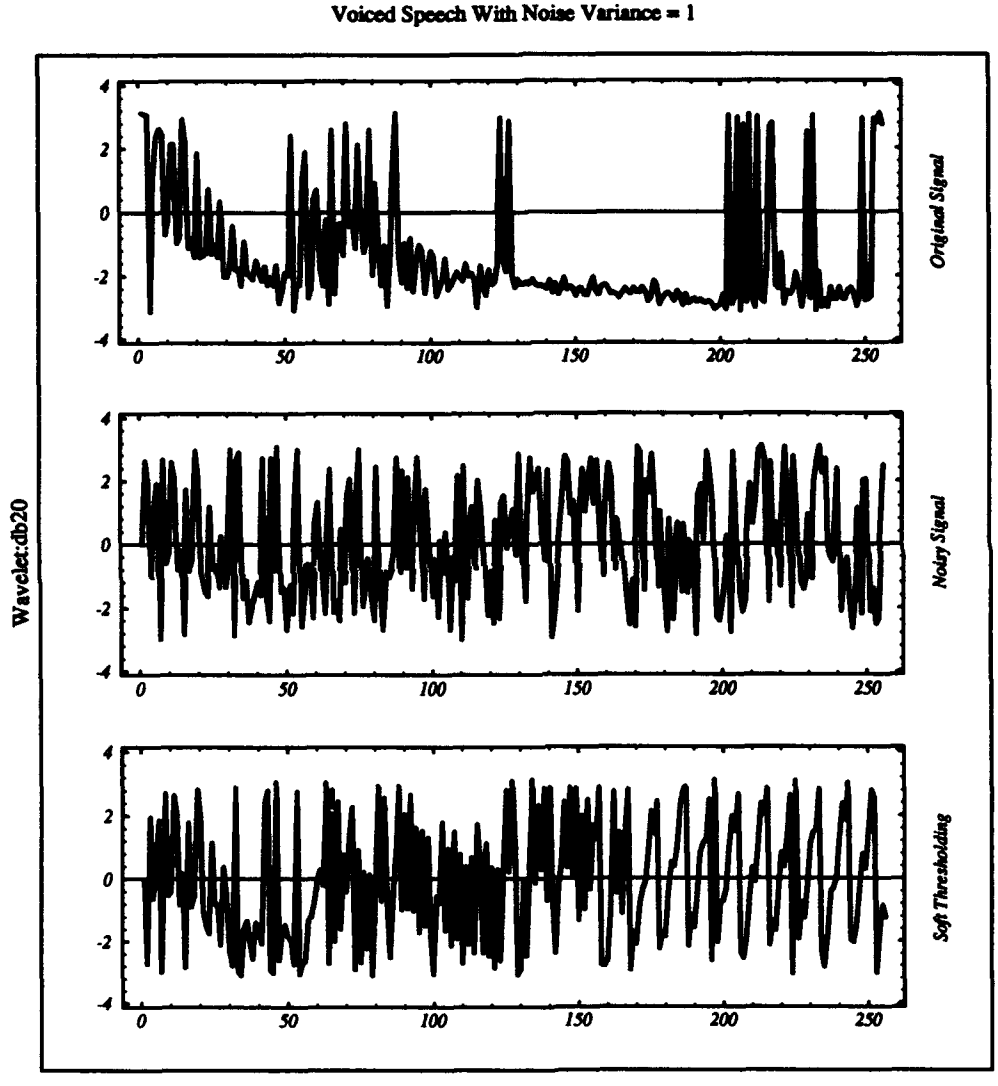

Figure E.9 Phase Of the FFT of the noisy voiced speech after the STT ($\sigma^2 = 1$).

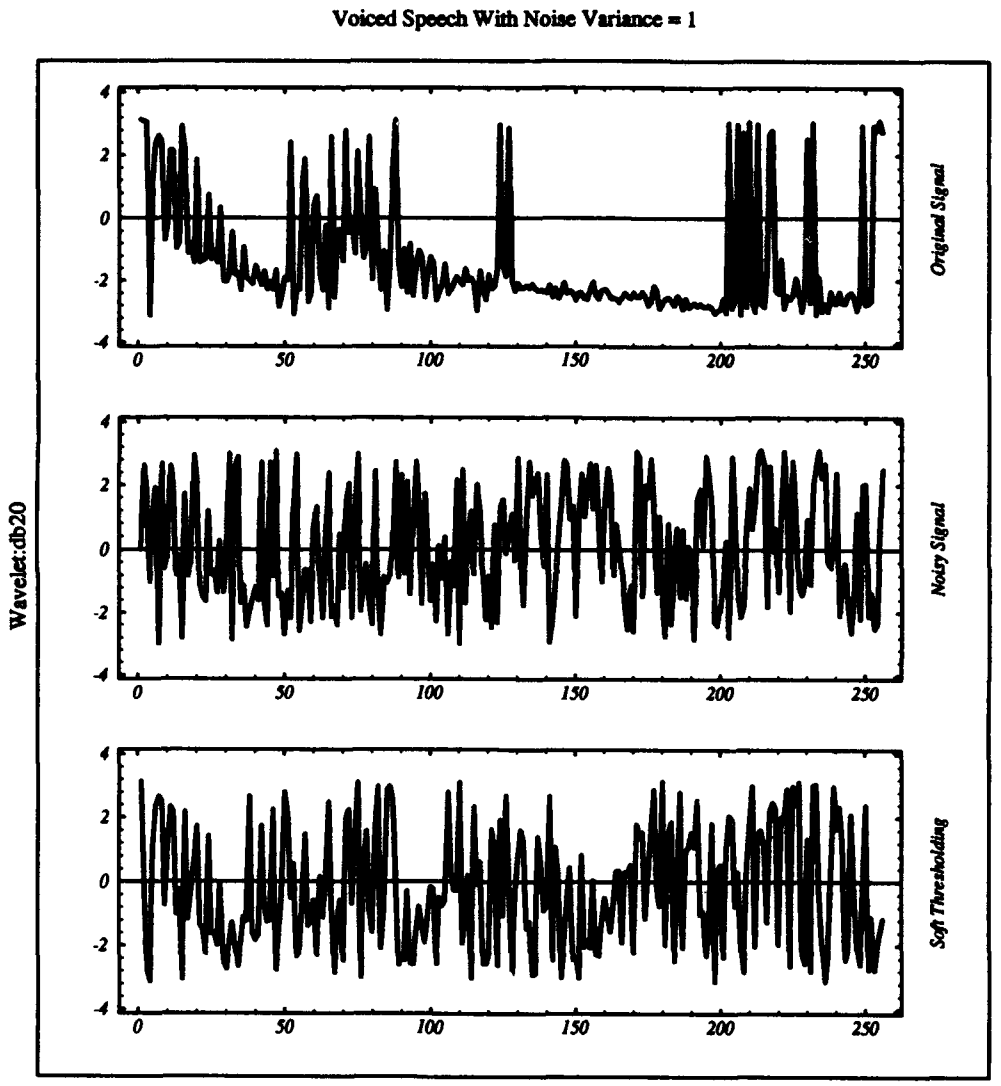

Figure E.10 Phase of the FFT of the clean voiced speech after the STT ($\sigma^2 = 1$).

Appendix F. Total Squared Error With Respect To Both The Clean And Noisy Speech Signals Using Compactly Supported Wavelets

This Appendix contains bar-charts showing the total squared error (TSE) between the de-noised speech signals and both the clean and noisy speech signal, using db6, coiflet(6), and db20. We studied the effects of both the soft and hard thresholding techniques (STT and HTT) on seven different noisy signals with signals-to-noise-ratios (SNR): -10db, -6db, -3db, 0db, 3db, 6db, and 10db. Eight different speech de-noising systems (SDS) have been studied:

1. WRINP means that the SDS uses the wavelet transform on the real and imaginary parts of the Fourier transform of the original noisy signal. This method reconstructs the signal using the phase of the original noisy speech signal.

2. WRI means that the SDS uses the wavelet transform on the real and imaginary parts of the Fourier transform of the original noisy signal. This method does not use the phase of the original noisy speech signal.

3. WTNP means that the SDS uses the wavelet transform of the original noisy signal (no Fourier transform). This method reconstructs the signal using the phase of the original noisy speech signal.

4. WT means that the SDS uses the wavelet transform of the original noisy signal (no Fourier transform). This method does not use the phase of the original noisy speech signal. This method is based on Donoho's original work on wavelet shrinkage.

5. SRINP means that the SDS uses Stein's criteria directly on the real and imaginary parts of the Fourier transform of the original noisy signal. This method reconstructs the signal using the phase of the original noisy speech signal. This method resembles the spectral subtraction developed by Steven Boll.

6. SRI means that the SDS uses Stein's criteria directly on the real and imaginary parts of the Fourier transform of the original noisy signal. This method does not use the phase of the

original noisy speech signal.

7. STNP means that the SDS uses Stein's criteria directly on the original noisy signal (no Fourier transform). This method reconstructs the signal using the phase of the original noisy speech signal.

8. ST means that the SDS uses Stein's criteria directly on the original noisy signal (no Fourier transform). This method does not use the phase of the original noisy speech signal.

The bar-charts (SNRs 0db and 6db), of the TSE with respect to the noisy signal (see figures F.1 and F.3), show the closeness between the de-noised signals and the noisy signal. Ideally, we would like the processed speech signals to be as far away as possible from the noisy signal, indicated by large TSE. All the bar-charts illustrate the fact that the STT outperforms the HTT, since all the STT bars have higher TSEs than the HTT bars.

The bar-charts (SNRs 0db and 6db), of the TSE with respect to the clean signal (see figures F.2 and F.4), show the closeness between the de-noised signals and the clean signal. Ideally, we would like the processed speech signals to be very close to the clean signal, indicated by small TSE. All the bar-charts illustrate the fact that the STT outperforms the HTT, since all the STT bars have lower TSEs than the HTT bars.

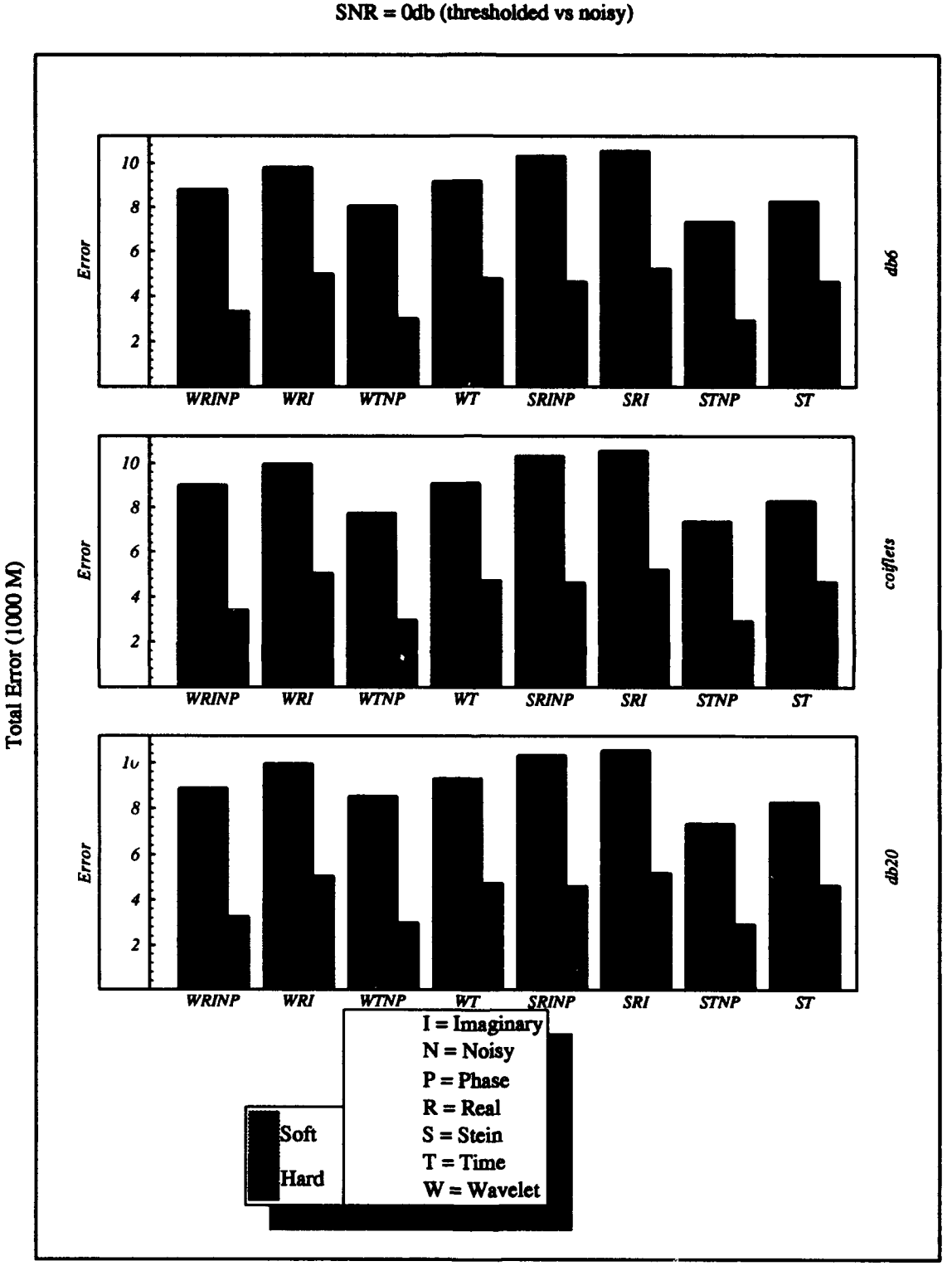

Figure F.1 TSE using noisy speech and the de-noised speech (0db) with wavelets: db6, coiflets, and db20.

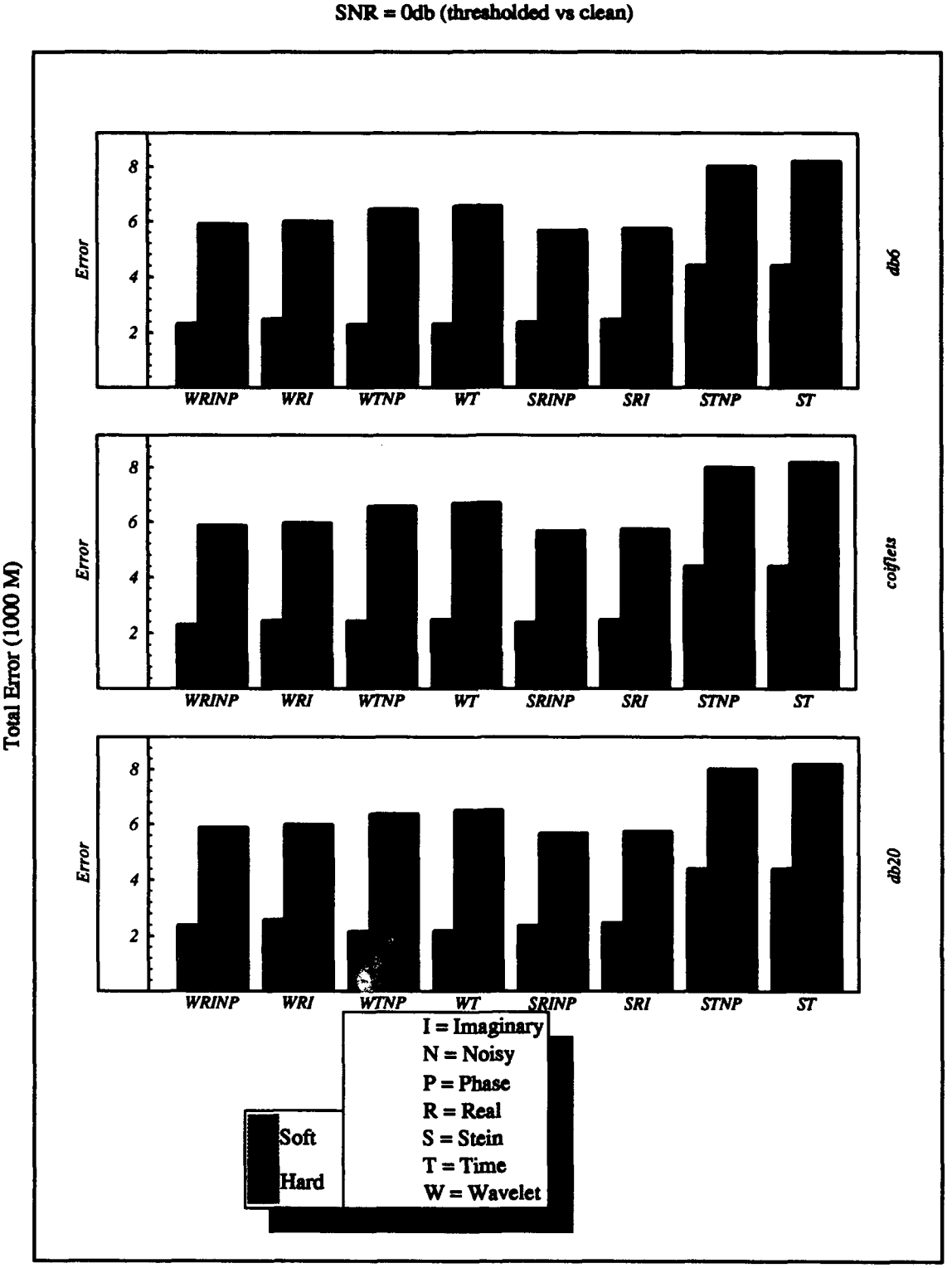

Figure F.2 TSE using clean speech and the de-noised speech (0db) with wavelets: db6, coiflets, and db20.

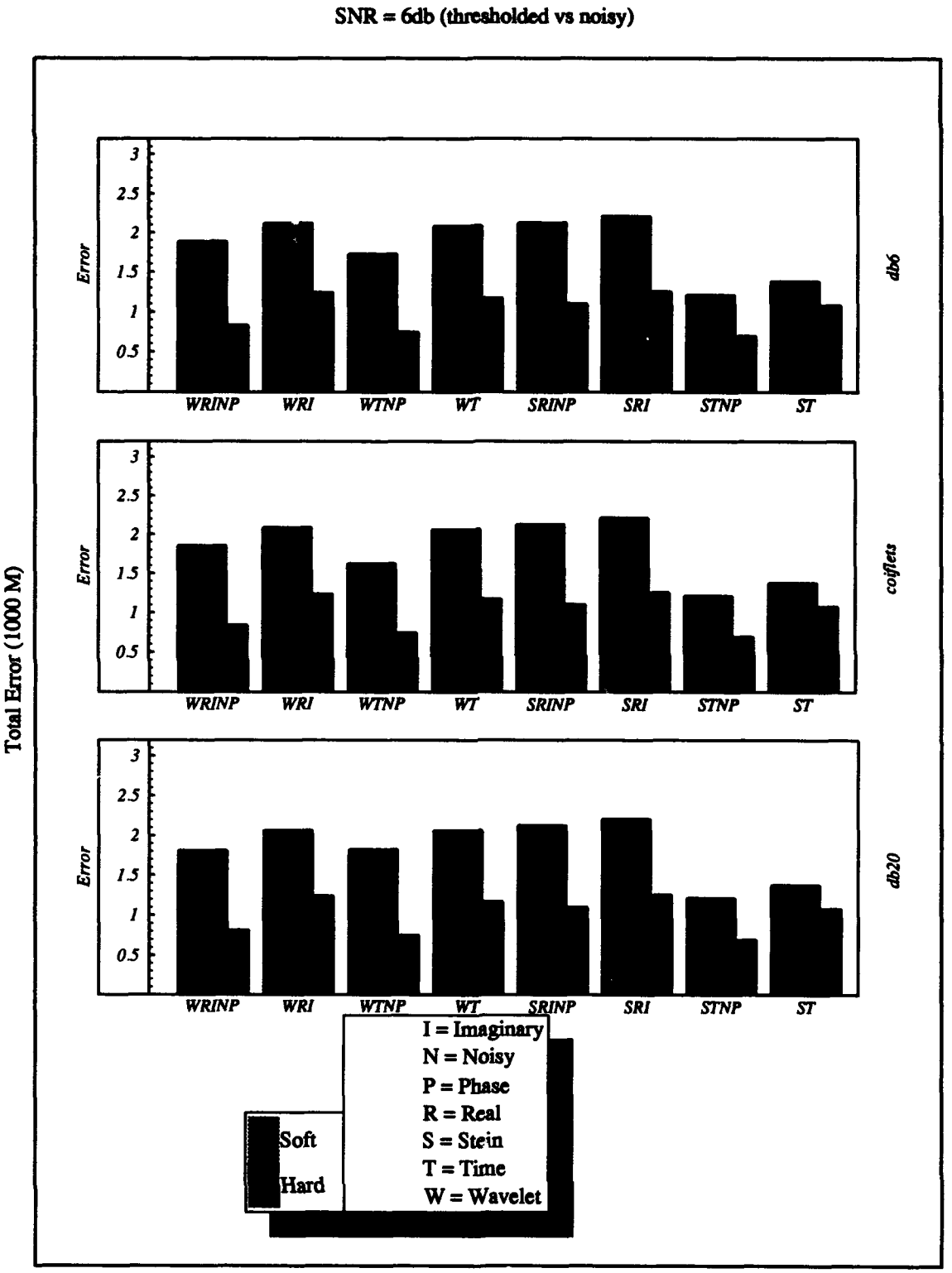

Figure F.3 TSE using the noisy speech and the de-noised speech (6db) with wavelets: db6, coiflets, and db20.

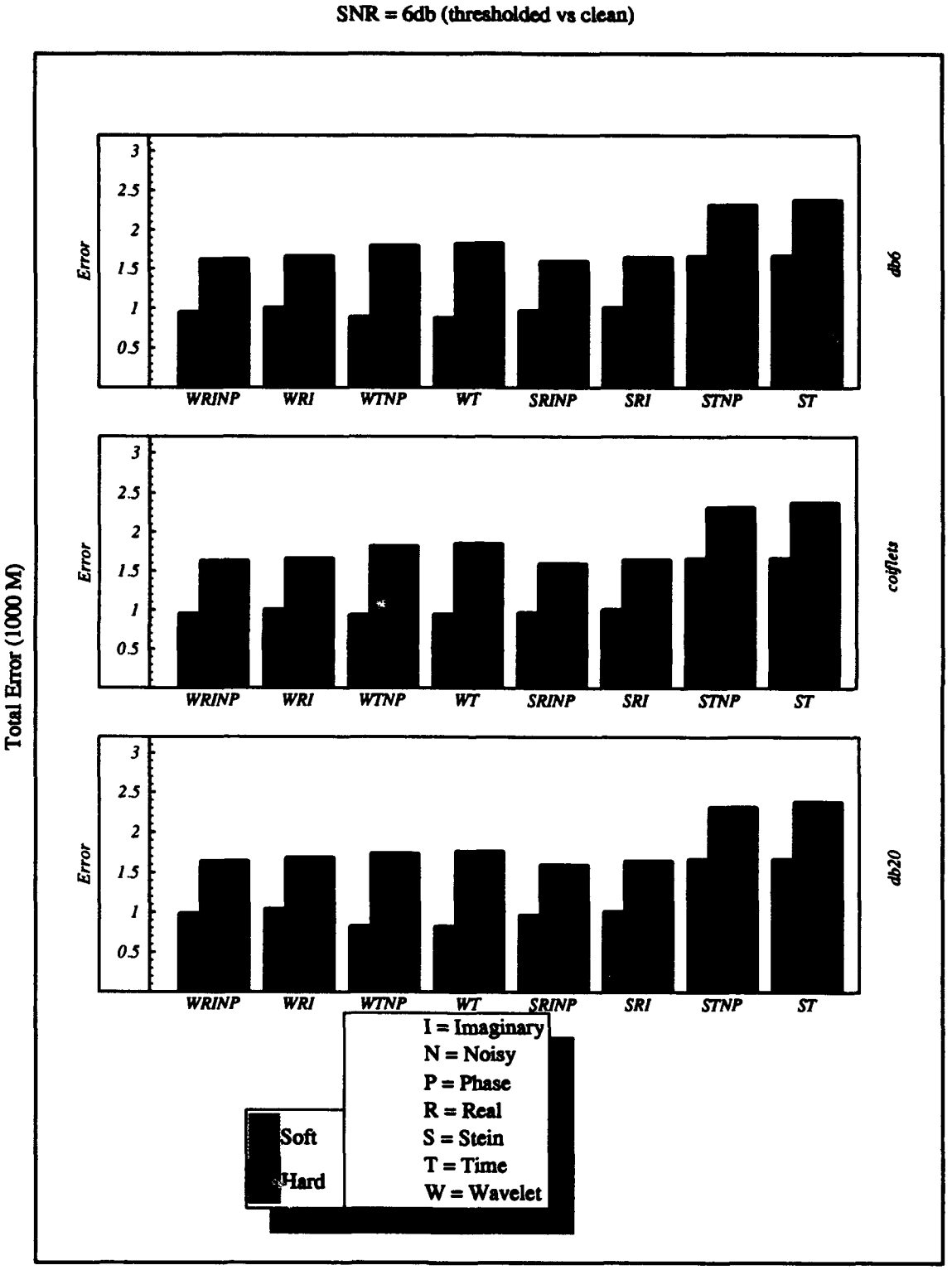

Figure F.4 TSE using the clean speech and the de-noised speech (6db) with wavelets: db6, coiflets, and db20.

Appendix G. Spectrum Analysis Of The Clean And Noisy Speech Signals

This appendix contains both the wide-band and narrow-band spectrograms of the clean speech signal, the 6db noisy speech signal, and the 0db noisy speech signal. Observe, the high energy of the first formant frequency (below 500Hz), the second and third formants frequencies (below 3kHz). In all figures, the vertical axis represents frequency and the horizontal axis represents samples of the signal (sampling frequency is 16kHz).

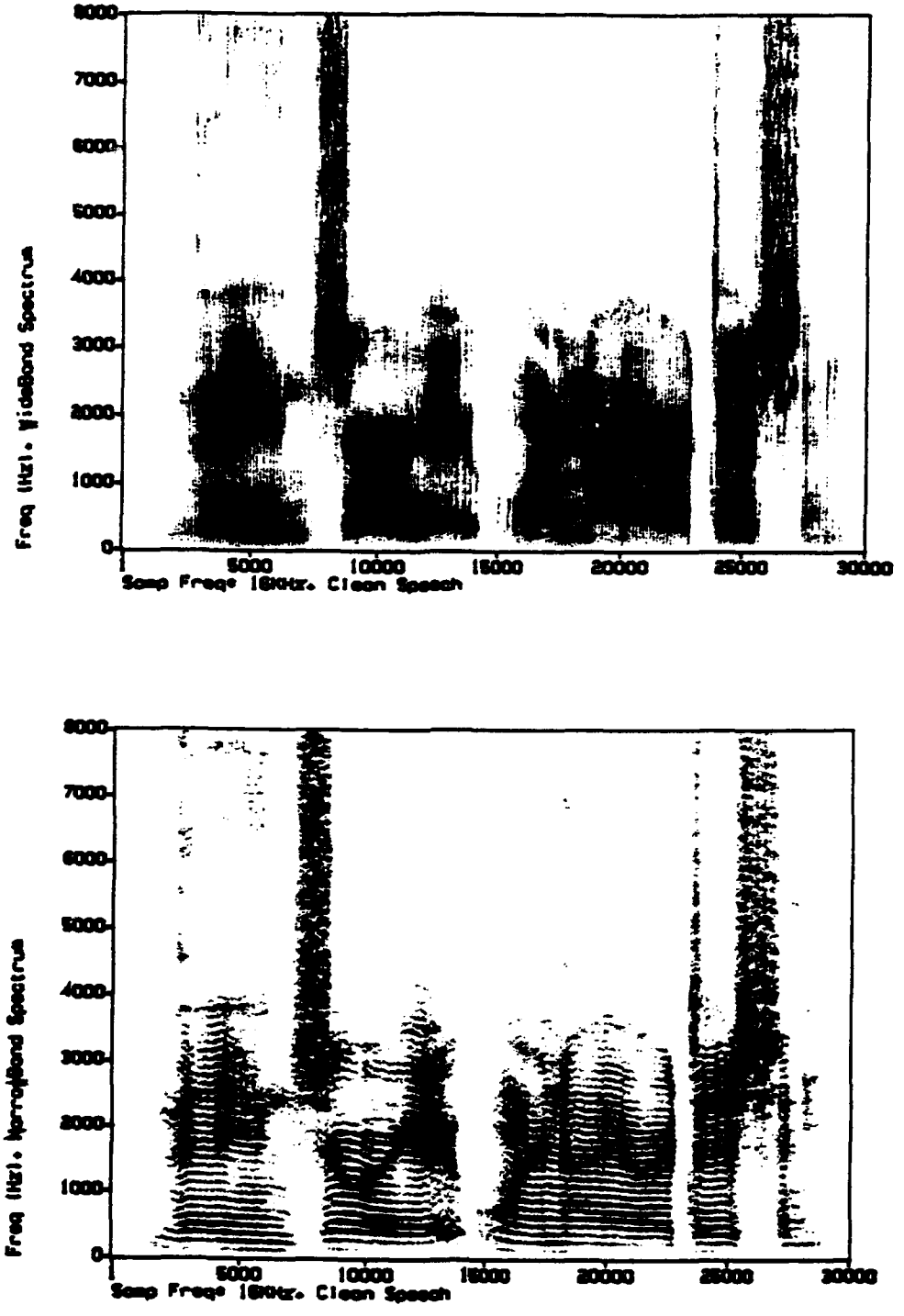

Figure G.1 Clean speech wide-band and narrow-band spectrums.

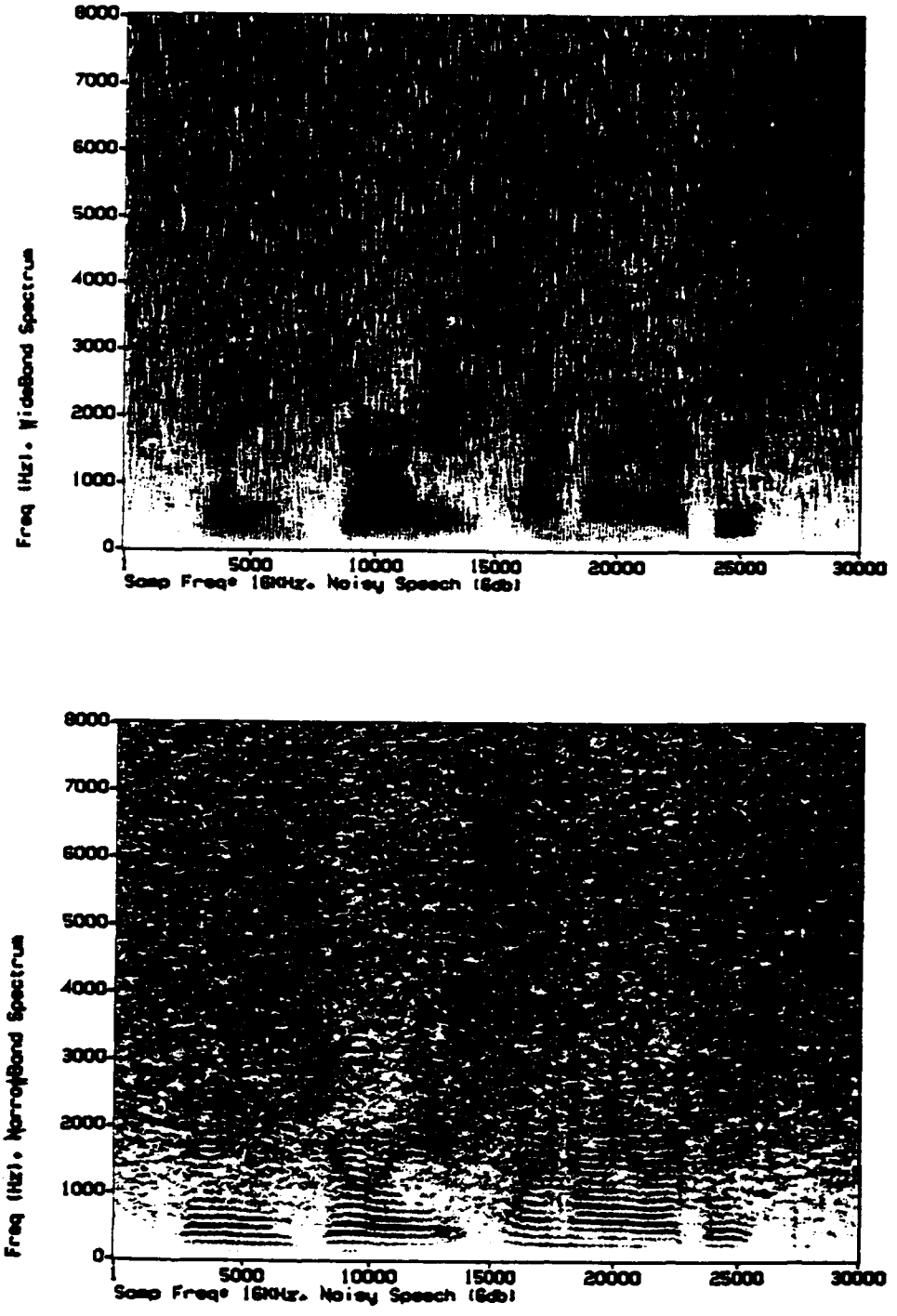

Figure G.2 Noisy speech wide-band and narrow-band spectrums (6db).

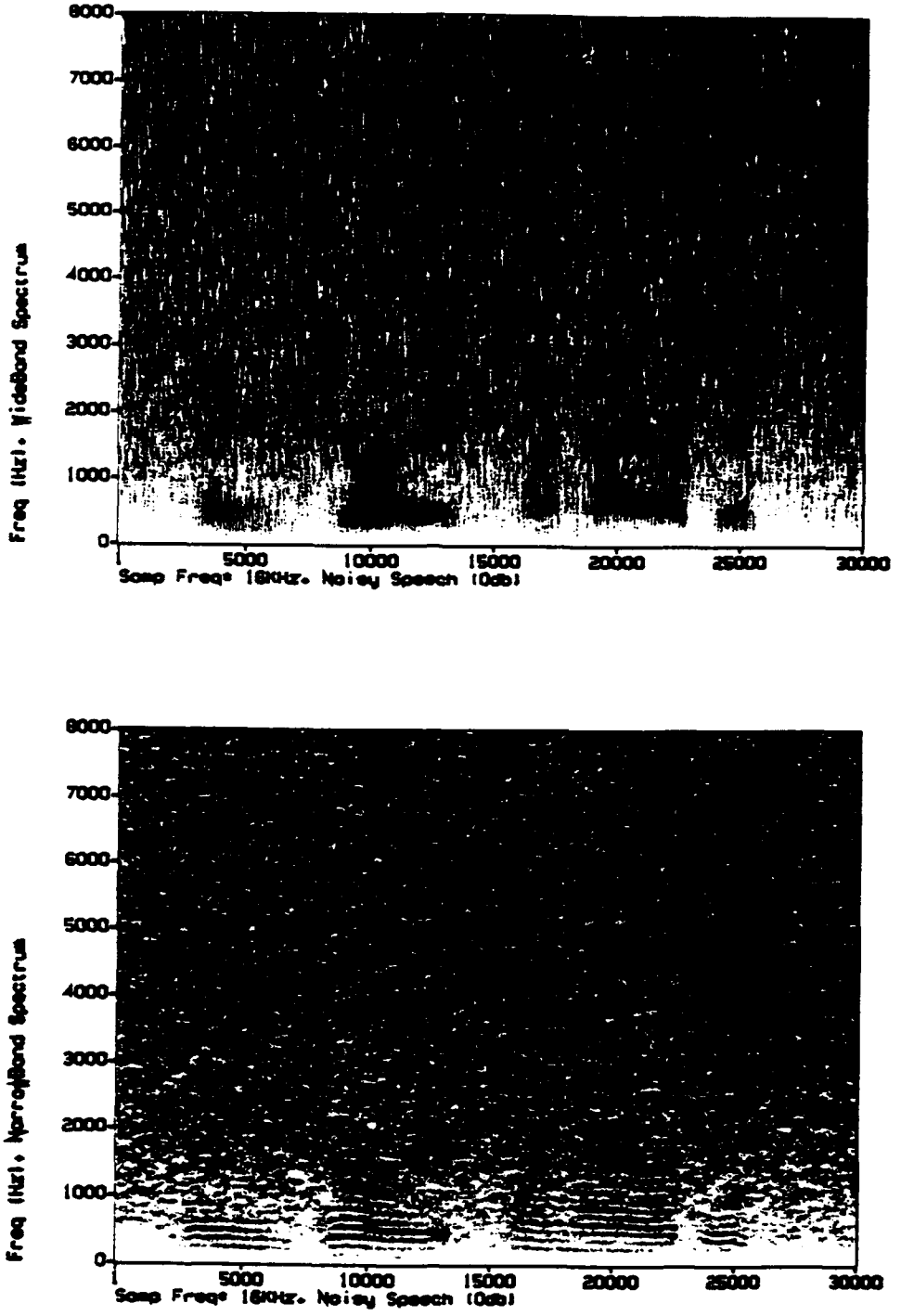

Figure G.3 Noisy speech wide-band and narrow-band spectrums (0db).

Appendix H. Spectrum Analysis Of The De-noised Speech Signals (0db and 6db) Without Using Wavelets

This appendix contains the wide-band spectrograms of two de-noised speech signals (0db and 6db). The speech signals were processed using the soft thresholding technique (STT) and Stein's criteria. Figure H.1 shows the noisy speech data processed in the time domain using Stein's criteria. Observe that when the signal-to-noise ratio (SNR) is 6db, all the formant frequencies are still preserved, however, the third formant frequency of the 0db processed speech signal was affected by the non-linear shrinkage. Figure H.2 shows the effects of applying the STT to the real and imaginary parts of the Fourier transform of the original signal. The original noisy phase was used before reconstruction of the de-noised speech signal. Observe that as the noise level increases (i.e., from 6db to 0db), the formants are affected. In all figures, the vertical axis represents frequency and the horizontal axis represents samples of the signal (sampling frequency is 16kHz).

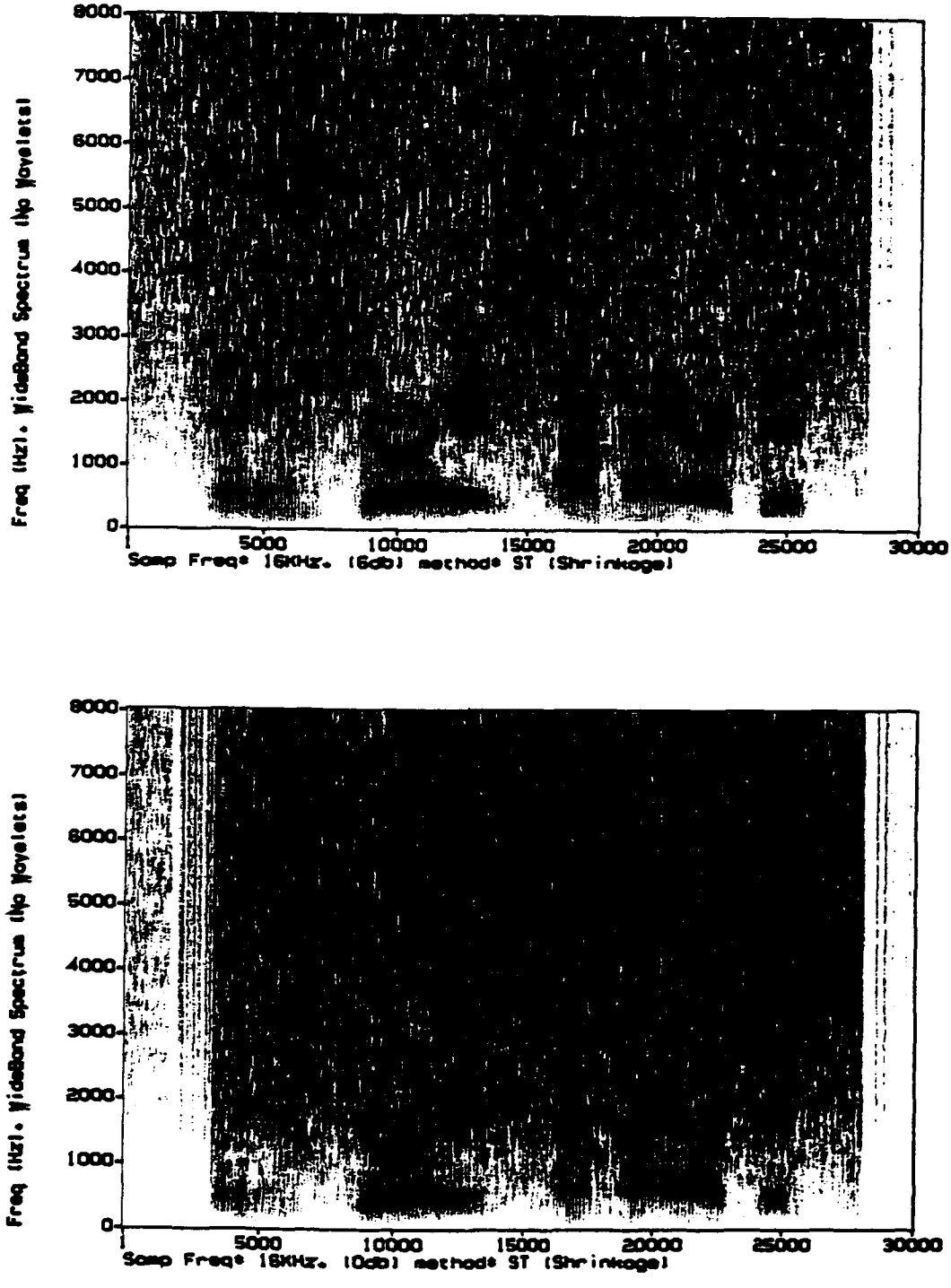

Figure H.1 De-noised speech using (ST) wide-band spectrum (0db and 6db).

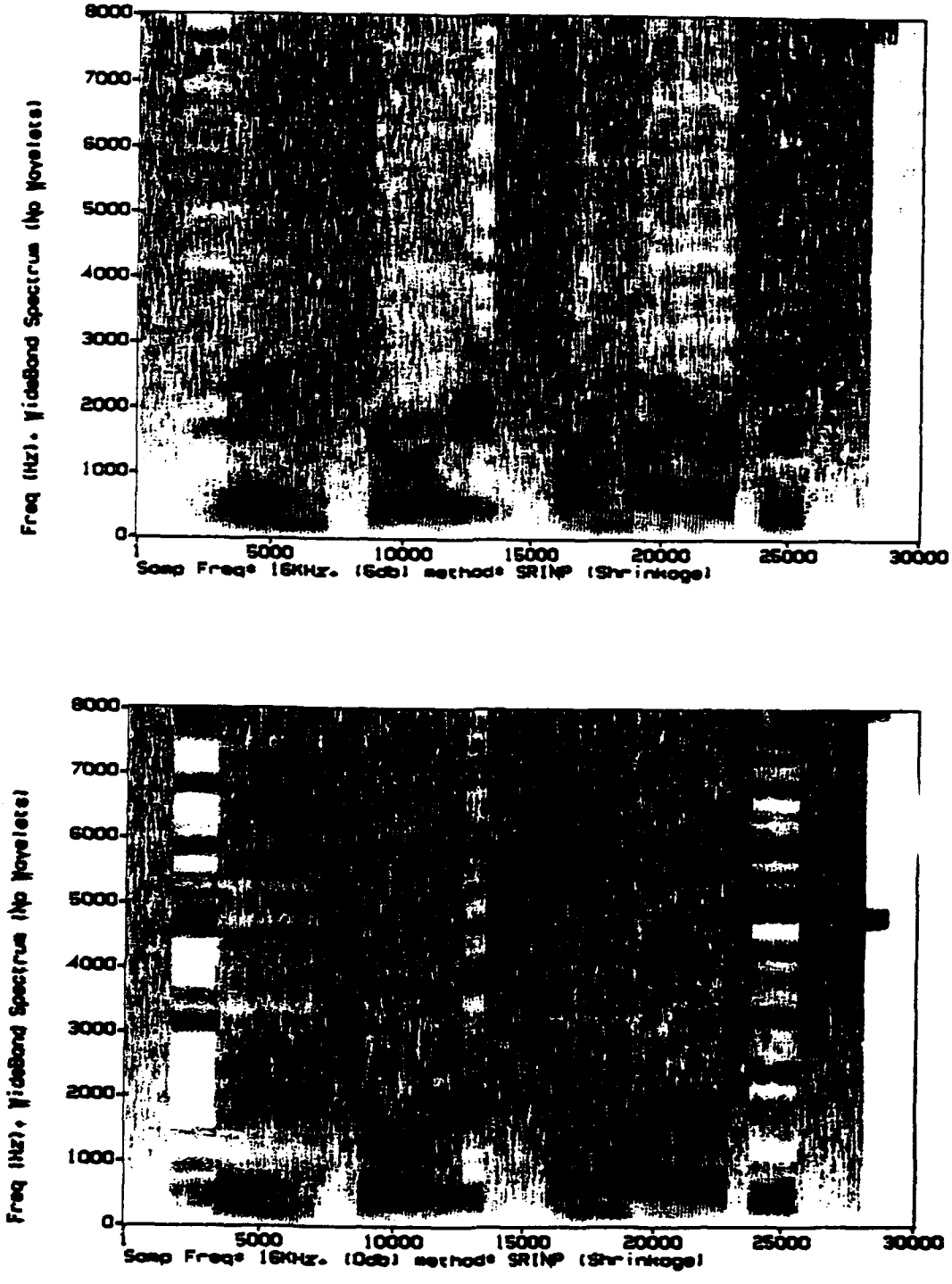

Figure H.2 De-noised speech using (SRINP) wide-band spectrums (0db and 6db).

Appendix I. Spectrum Analysis Of The De-noised Speech Signals (0db and 6db) Using Wavelets In Time

This appendix contains the wide-band spectrograms of two de-noised speech signals (0db and 6db). The speech signals were processed using the soft thresholding technique (STT) and Stein's criteria was applied to the wavelet transform. Observe the aliasing produced by db6 and coiflet(6). This aliasing is mainly due to the fact that these wavelets have Fourier transforms with many high energy side-lobes. All wavelets used (db6, coiflet(6), and db20) preserve most of the formant frequencies. Notice the performance of the db20; no aliasing and very clear formant frequencies. In all figures, the vertical axis represents frequency and the horizontal axis represents samples of the signal (sampling frequency is 16kHz).

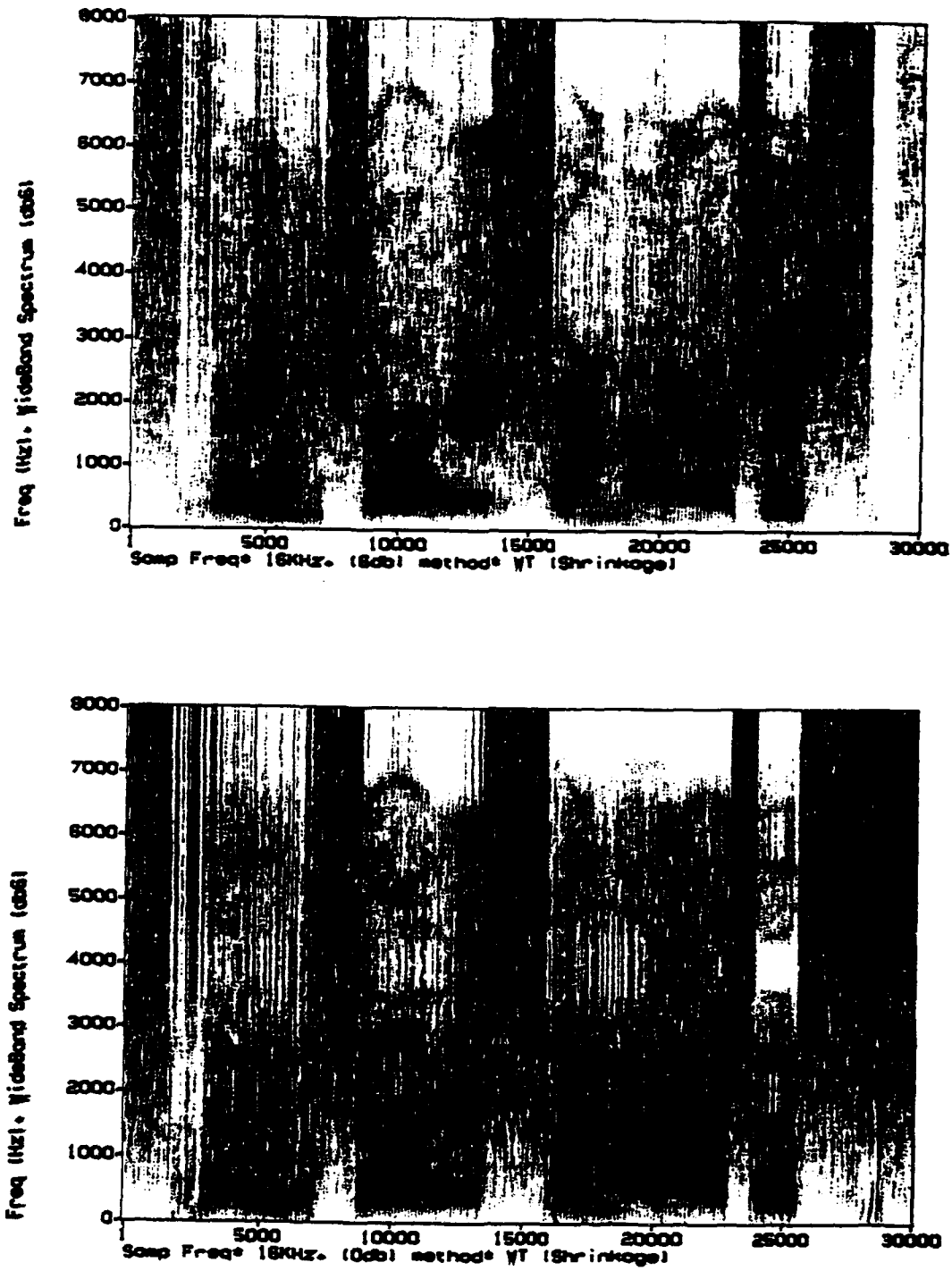

Figure I.1 De-noised speech using (WT and wavelet db6) wide-band spectrums (0db and 6db).

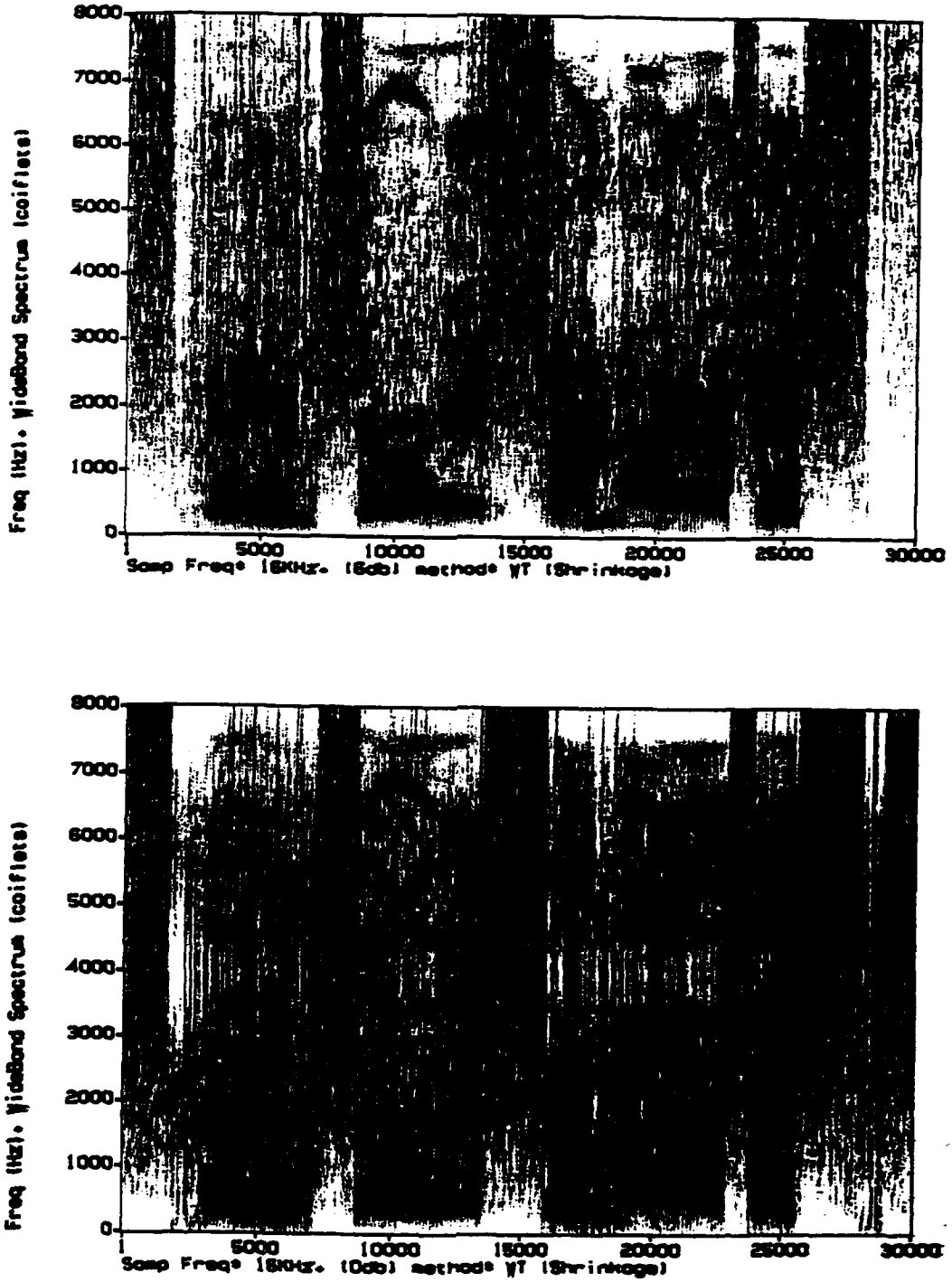

Figure I.2 De-noised speech using (WT and wavelet coiflet(6)) wide-band spectrums (0db and 6db).

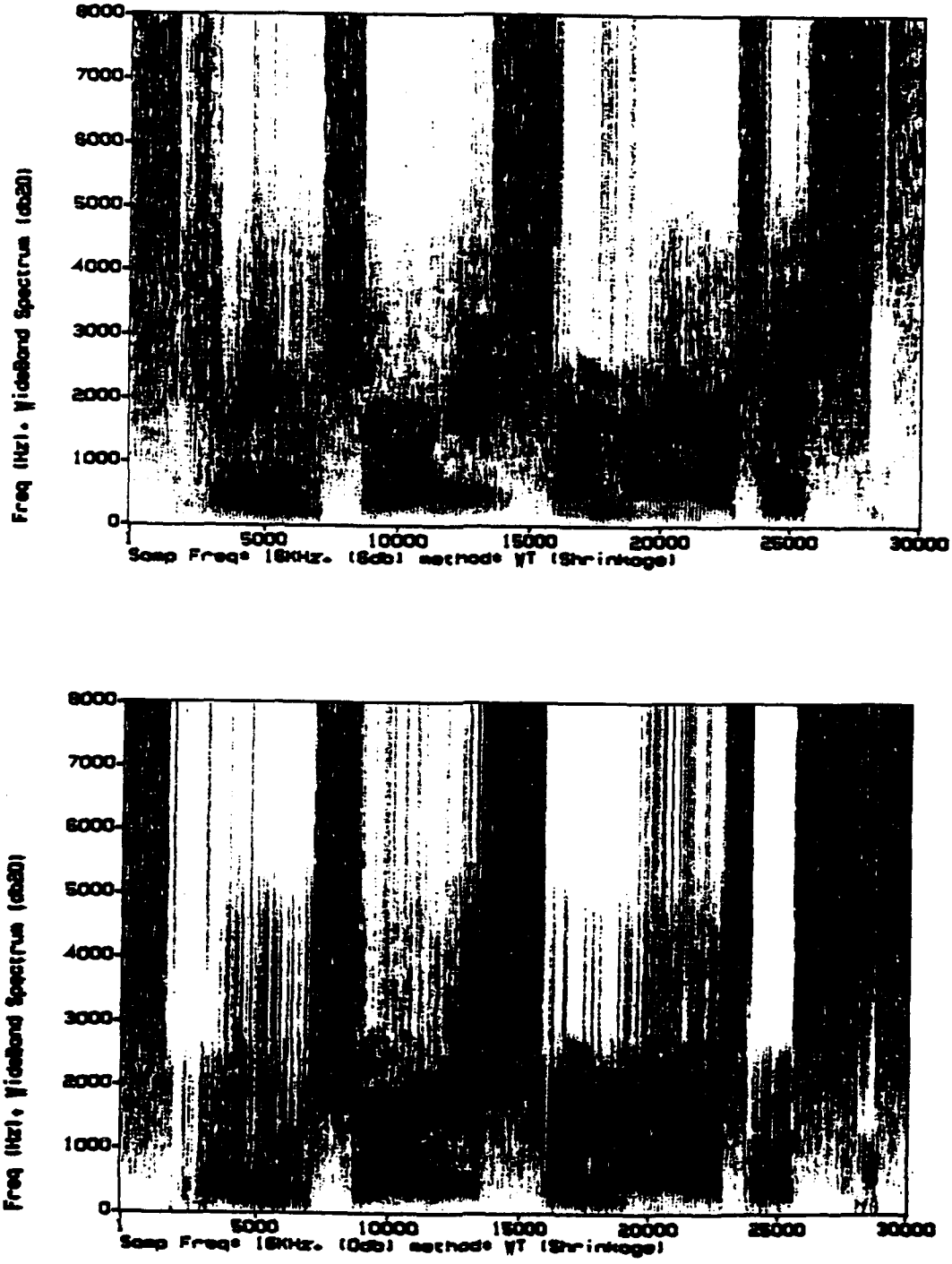

Figure I.3 De-noised speech using (WT and wavelet db20) wide-band spectrums (0db and 6db).

Appendix J. Spectrum Analysis Of The De-noised Speech Signals (0db and 6db) Using Wavelets In Fourier

This appendix contains the wide-band spectrograms of two de-noised speech signals (0db and 6db). The speech signals were processed using the soft thresholding technique (STT) and Stein's criteria was applied to the wavelet transforms of both the real and imaginary parts of the Fourier transform of the original noisy signals. The original noisy phase was used before reconstruction of the de-noised signal. Observe that most of the formant frequencies are still preserved for all wavelets used and that there is no noticeable aliasing caused by any wavelet. In all figures, the vertical axis represents frequency and the horizontal axis represents samples of the signal (sampling frequency is 16kHz).

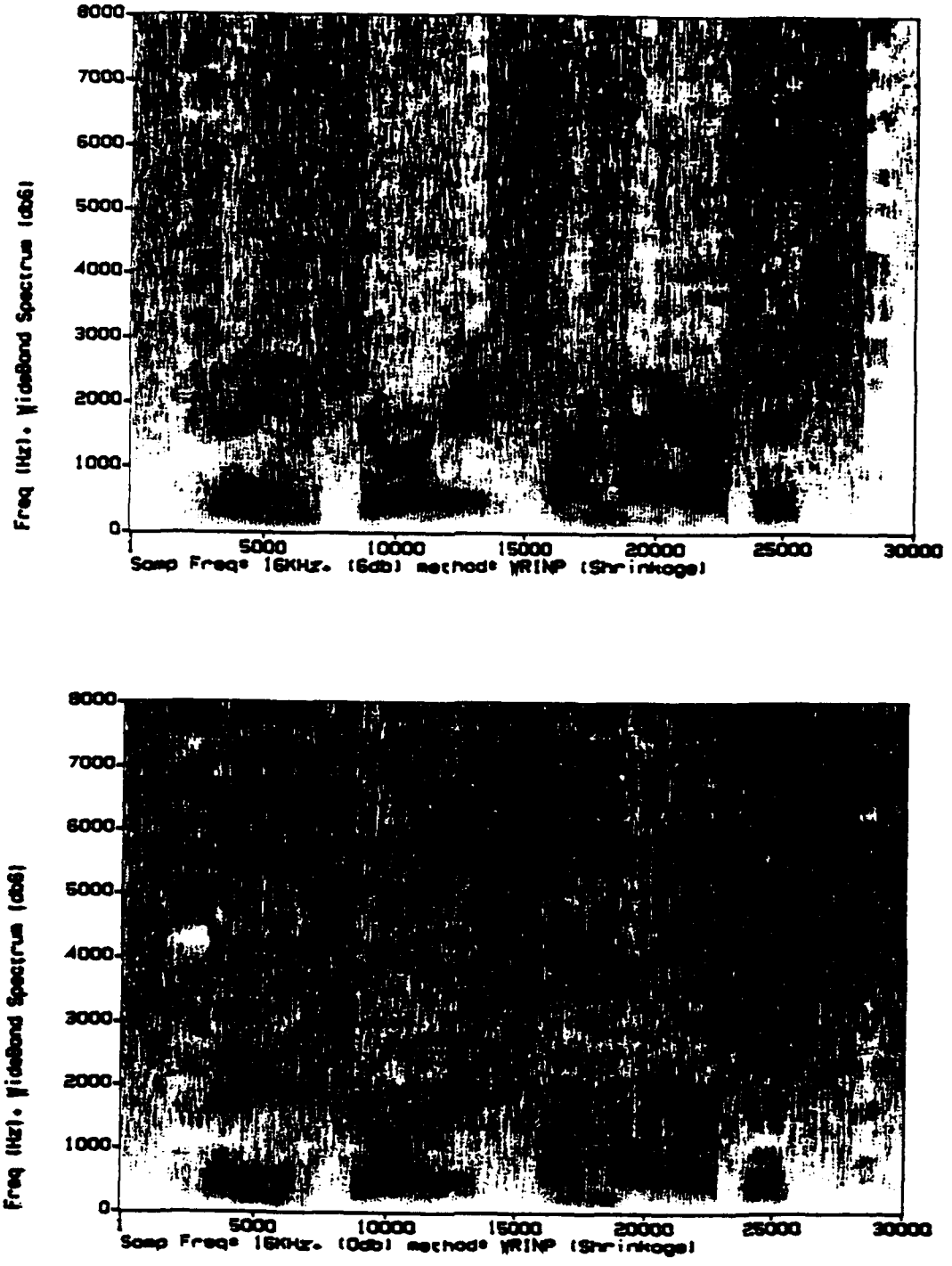

Figure J.1 De-noised speech using (WRINP and wavelet db6) wide-band spectrums (0db and 6db).

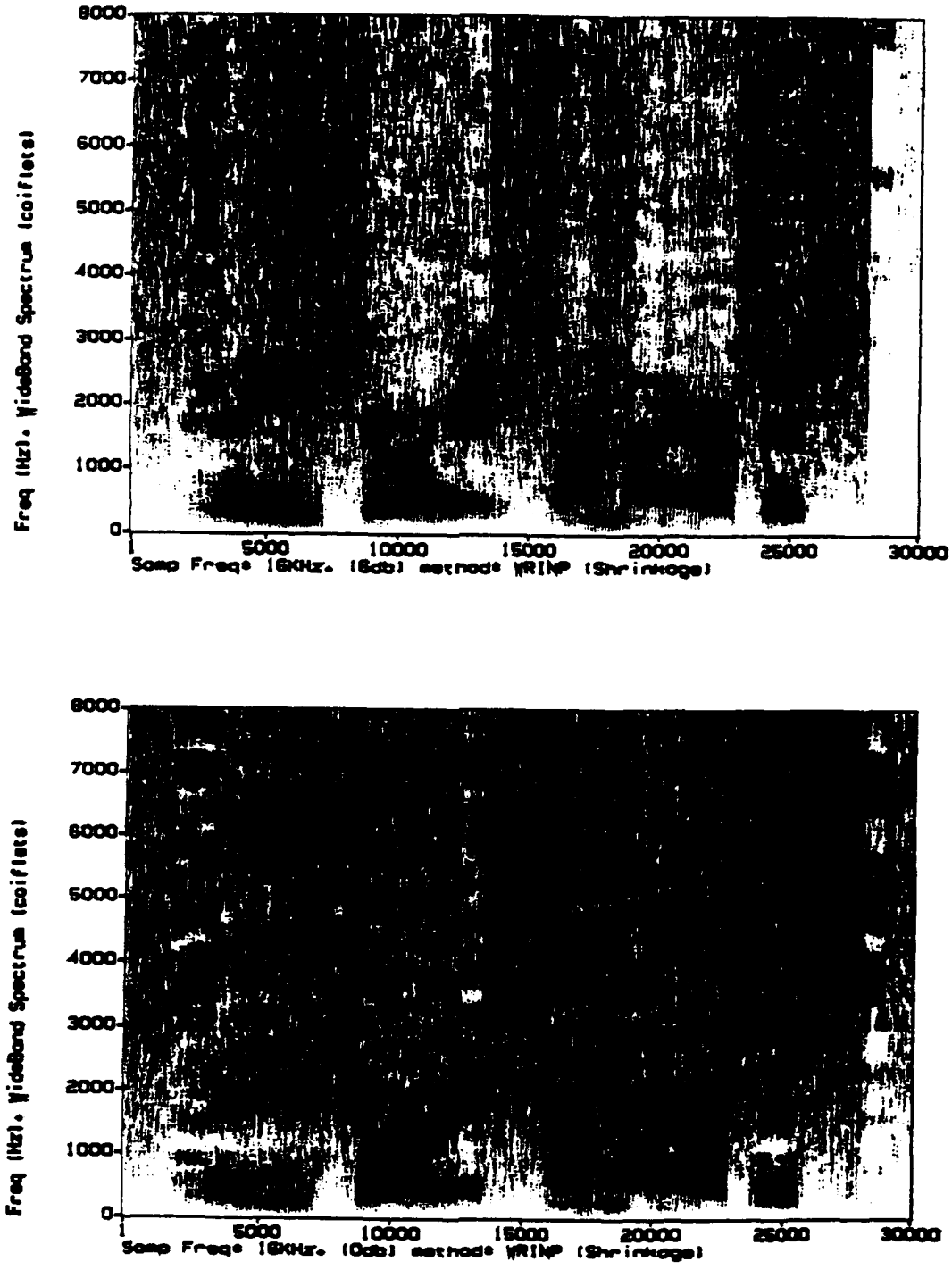

Figure J.2 De-noised speech using (WRINP and wavelet coiflet(6)) wide-band spectrums (0db and 6db).

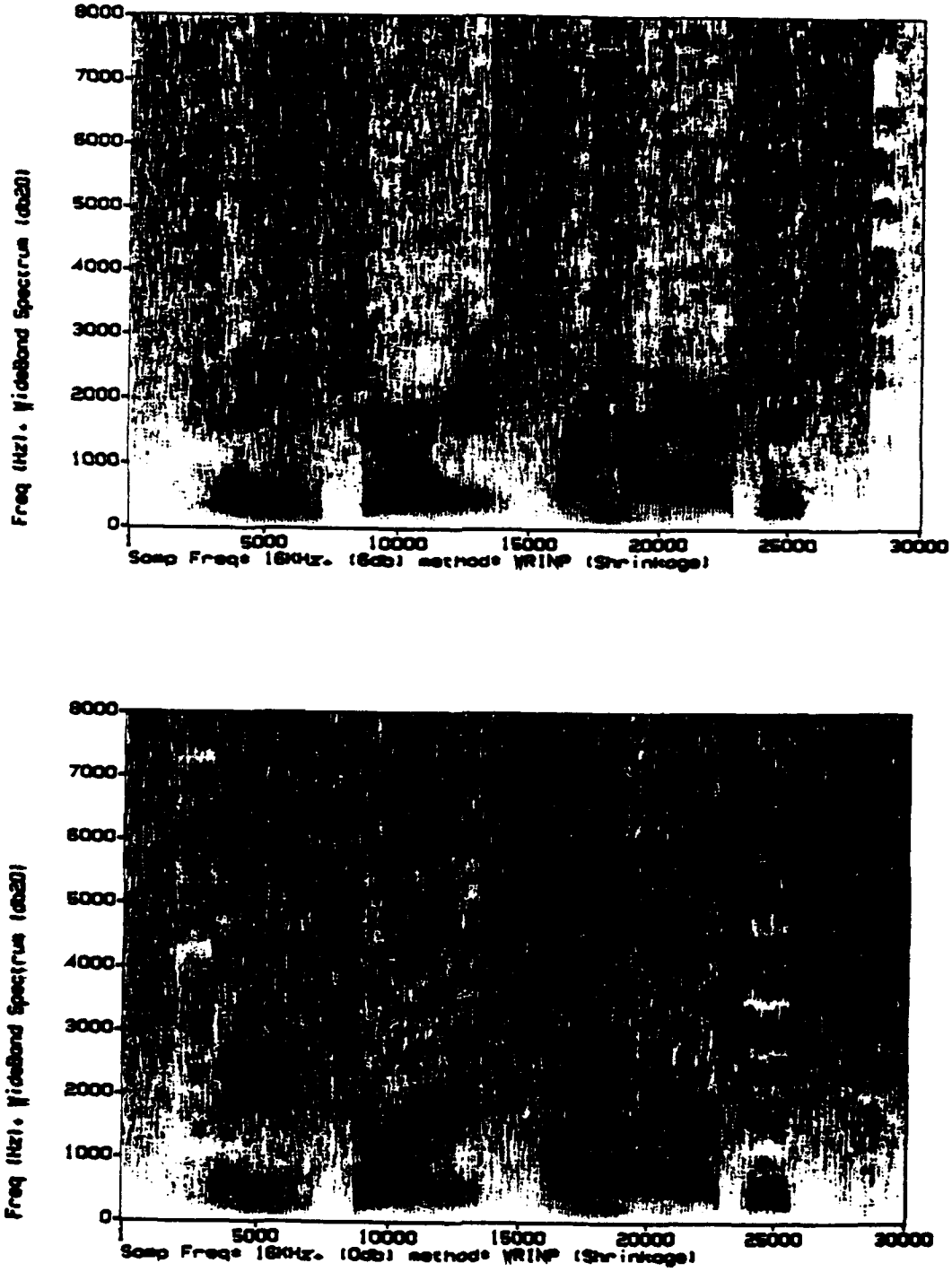

Figure J.3 De-noised speech using (WRINP and wavelet db20) wide-band spectrums (0db and 6db).

Bibliography

1. Anderson, B.P. ***Theory and Implementation of Wavelet Analyses in Rational Resolution Decompositions***. MS thesis, Air Force Institute of Technology, December 1992.
2. Apostol, M.T. ***Mathematical Analysis*** (second Edition). Addison Wesley Publishing company, 1974.
3. Boll, S.F. "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," ***IEEE Transactions on Acoustics, Speech, and Signal Processing***, *ASSP-27*(2):113–120 (1979).
4. Daubechies, I. ***Ten Lectures on Wavelets***. Society for Industrial and Applied Mathematics, 1992.
5. Donoho, D.L. ***De-noising by Soft-thresholding***. Technical Report, Stanford University, 1992.
6. Donoho, D.L. ***Ideal Spatial Adaptation by Wavelet Shrinkage***. Technical Report, Stanford University, 1992.
7. Donoho, D.L. ***Wavelet Shrinkage and W.V.D.: A 10-Minute Tour***. Technical Report, Stanford University, 1992.
8. Donoho, D.L. and I.M. Johnstone. ***Adapting to Unknown Smoothness via Wavelet Shrinkage***. Technical Report, Stanford University, 1991.
9. Donoho, D.L. and I.M. Johnstone. ***Minimax Estimation via Wavelet Shrinkage***. Technical Report, Stanford University, 1991.
10. Etter, W., *et al.* "Adaptive Noise Reduction Using Discrimination Functions," ***IEEE*** (1991).
11. Gurgen, F.S. and C.S. Chen. "Speech Enhancement by Fourier-Bessel Coefficients fo Speech and Noise," *IEE Proceedings*, *137*(5):290–294 (1990).
12. Hogg, R.V. and A.T. Craig. ***Introduction to Mathematical Statistics*** (Fourth Edition). Macmillan Publishing Co., Inc., 1978.
13. Kabrisky, M., *et al.* "Reconstruction of Mutilated Speech," ***IEEE AES Magazine*** (1989).
14. Kay, S.M. ***Modern Spectral Estimation: Theory and Application***. Signal Processing Series, Prentice-Hall, 1988.
15. Kobatakeand, H., *et al.* "Enhancement of Noisy Speech by Maximum Likelihood Estimation," ***IEEE*** (1991).
16. Mallat, S.G. and S. Zhong. ***Complete Signal Representation with Multiscale Edges***. Technical Report, New York University, 1989.
17. Oppenheim, A.V. and R.W. Schafer. ***Discrete-Time Signal Processing***. Prentice-Hall, 1989.
18. Papoulis, A. ***Probability, Random Variables, and Stochastic Processes*** (Third Edition). McGraw-Hill, Inc., 1991.
19. Parsons, T.W. ***Voice and Speech Processing***. McGraw-Hill, Inc., 1987.
20. Stein, C.M. "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, *9*(6):1135–1151 (1981).
21. Warhola, G., *et al.* "Applications of Wavelets to Signal Processing,". AFIT Science And Research Center, 1991.
22. Watson, G.N. ***A treatise on the theory of Bessel functions*** (2nd Edition). Cambridge University Press, 1966.

Vita

Lieutenant Hassan Dehmani was born Dec 27, 1968 in Casablanca, Morocco. At the age of sixteen, he attended the College Royal Preparatoire aux Techniques Aeronautiques (CRPTA) in the southern city of Marrakech, Morocco. Upon graduation in the summer of 1988, with a high school diploma in mathematics, Lieutenant Dehmani received a scholarship to study in the United States of America. He was selected to continue his education at the United States Air Force Academy (USAFA) in Colorado Springs, Colorado. He graduated with a Bachelor of Science in Electrical Engineering with a strong emphasis on computer engineering and mathematics.

Upon graduation from USAFA in the summer of 1992, Lieutenant Dehmani received a second scholarship to pursue a Master's Degree in Computer Engineering with an emphasis on Digital Signal Processing at the United States Air Force Institute Of Technology.

Lieutenant Dehmani married Danielle A. Dix on Valentine's Day 14 February 1994. Upon graduation from AFIT, he will be assigned to the First Air Base in the city of Salé, Morocco. There he will be in charge of the Computer Center of the First Air Base.

Permanent address: 65 Rue Rabia-Al-Adaouiya
Casa03
Casablanca, Morocco

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	JUNE, 1994	MASTER'S THESIS

4. TITLE AND SUBTITLE
NOISE REDUCTION FOR SPEECH ENHANCEMENT USING NON-LINEAR WAVELET PROCESSING

5. FUNDING NUMBERS

6. AUTHOR(S)
HASSAN DEHMANI, LIEUTENANT, ROYAL MOROCCAN AIR FORCE (RMAF)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
AIR FORCE INSTITUTE OF TECHNOLOGY, WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/GCE/ENC/94J-1

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
DEPARTMENT OF DEFENSE
R52
FT MEADE, MD 20755-6000
CONTRACT NUMBER: H98230-R5-93-9187

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*
THE PROBLEM OF SPEECH ENHANCEMENT PRESENTS MANY OBSTACLES IN THE SPEECH PROCESSING FIELD. THIS THESIS DEVELOPS SEVERAL SPEECH DE-NOISING SYSTEMS THAT CAN BE USED IN THE TIME, FOURIER, AND WAVELET DOMAINS. WE PRESENT TWO THRESHOLDING TECHNIQUES: SOFT AND HARD. THE APPLICATION OF THESE THRESHOLDING TECHNIQUES TO NOISY SPEECH DATA IS DISCUSSED. THE COMBINATION OF BOTH WAVELETS AND THE FOURIER DOMAINS WITH NOISY PHASE RESTORATION PROVES TO YIELD THE BEST RESULTS IN TERMS OF INTELLIGIBITY. INFORMAL LISTENING TESTS WERE CONDUCTED IN ORDER TO COMPARE THE EFFECTS AND DIFFERENCES BETWEEN THE SPEECH DE-NOISING SYSTEMS.

14. SUBJECT TERMS
WAVELETS, SPEECH, MULTIRESOLUTION ANALYSIS, NOISE REDUCTION, SHRINKAGE

15. NUMBER OF PAGES
199

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL